How Do Causal Structures Affect Model Selection? Heterogeneous Treatment Effect Estimation with Observational Data

Juho Eerik Lähteenmaa
University of Helsinki
Faculty of Social Sciences
Economics
Master's Thesis
May 2020

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta – Fakultet – Faculty Faculty of Social Sciences		Koulutusohjelma – Utbildingsprogram – Degree Programme Department of Political and Economic Studies
Tekijä – Författare – Author Juho Lähteenmaa		
Työn nimi – Arbetets titel – Title How Do Causal Structures Affect Model Selection? Heterogeneous Treatment Effect Estimation with Observational Data		
Oppiaine/Opintosuunta – Läroämne/Studieinriktning – Subject/Study track Economics		
Työn laji – Arbetets art – Level Master's Thesis	Aika – Datum – Month and year May 2020	Sivumäärä – Sidoantal – Number of pages 71 (77 with Appendices)

Tiivistelmä – Referat – Abstract

In social sciences, as in health sciences, there is an increasing interest in exploring differences in treatment effects amongst subpopulations and even individuals. In many cases, researchers must rely on observational data where the assignment mechanism of the treatment is non-randomized. Nevertheless, by including a sufficient set of covariates in the used model, it is possible to draw a causal inference. However, some causal structures have proved to cause bias in the treatment effect estimates when particular pre-treated variables in them are conditioned. In existing literature there is no consensus as to how to treat these structures, especially in the heterogeneous treatment effect estimation case.

The aim of this thesis is to explore how causal structures affect covariate selection in the heterogeneous treatment effect estimation context. The theoretical background of this subject is built on the potential outcomes framework and structural causal models. This thesis provides an overview of heterogeneous treatment effect estimation methods, including a more detailed view on the causal forest method. The second stage of the thesis is carried out by executing a simulation study where the causal forest method is applied with different causal structures. In each simulation, different sets of conditioned covariates are tested.

The simulation study results prove almost consistent. In every simulation except one, a higher number of variables implicates improvement in performance. Surprisingly, this result is applicable even to the cases where structural causal models literature suggests not to condition all the variables. According to the results of the simulation study, a practical recommendation would be to include as many relevant pre-treated, non-instrumental variables in the model as possible. The results are in line with practical recommendations given in potential outcomes framework literature.

Avainsanat – Nyckelord – Keywords
Heterogeneous treatment effects, observational data, structural causal models, potential outcomes

Muita tietoja – Övriga uppgifter – Additional information
The thesis is made for Hospital District of Helsinki and Uusimaa (HUS).

Contents

1 Introduction **3**

2 Review of Literature **5**

2.1 The Framework of Potential Outcomes 6
2.2 Treatment Effect and the Framework of Potential Outcomes . . 6
2.3 Average Treatment Effect 7
2.4 Randomized Controlled Trials 8
2.5 Observational Data and Unconfoundedness 9
2.6 Methods to Remove Selection Bias from Observational Data . . . 10
2.7 Balancing Scores and the Propensity Score 11
2.8 Limitations of the Potential Outcome Approach 13
2.9 Structural Causal Model 13
2.10 Basic Terminology for Graphical Causal Models 14
2.11 Directed Acyclic Graphs 15
2.12 The d-Separation Criterion 16
2.13 Interventions in DAGs 17
2.14 The Back-Door Criterion 19
2.15 Limitations of the SCM Approach 20
2.16 When to Adjust? 21
2.17 Heterogeneous Treatment Effects 24
2.18 An Overview of Treatment Effect Heterogeneity Estimation Methods 24
2.19 Regression Tree 27
2.20 Causal Tree 30
2.21 Causal Forest 31
2.22 Earlier Treatment Effect Estimation Simulation Studies 32
2.23 Simulation Study 34
2.23.1 Randomized Controlled Trial with Constant Treatment Effect, Including Covariates Affecting the Output 35
2.23.2 Constant Treatment Effect with Unconfounded Assignment, Including Covariates Affecting the Output 36
2.23.3 Constant Treatment Effect with Confounded Assignment, Including Covariates Affecting the Output 39
2.23.4 Constant Treatment Effect with Confounded Assignment, Including Covariates Affecting the Output and a Pure Local M-Structure 41
2.23.5 Constant Treatment Effect with Confounded Assignment, Including Covariates Affecting the Output and an Impure Local M-Structure 44
2.23.6 Randomized Controlled Trial with Heterogeneous Treatment Effect, Including Covariates Affecting the Output . 45
2.23.7 Heterogeneous Treatment Effect with Confounded Assignment, Including Covariates Affecting the Output 49

2.23.8 Heterogeneous Treatment Effect with Confounded Assignment, Including Covariates Affecting the Output and a Pure Local M-Structure 50
2.23.9 Heterogeneous Treatment Effect with Confounded Assignment, Including Covariates Affecting the Output and an Impure Local M-Structure 53

3 Discussion **56**

4 Conclusion **60**

A Unbiasedness of the ATE Estimator **71**

B Variance of the ATE Estimator **72**

C Unconfoundedness Given a Balancing Score **73**

D Balancing Property of the Propensity Score **73**

E A Simple Algorithm for Estimating the Propensity Score **74**

F Rules of do-calculus **74**

G K-Fold Cross-Validation Algorithm **75**

H Recursive Binary Split **76**

I Boosting for Regression Trees **76**

J Coefficients in the Simulation Study **77**

K R-Codes and used packages **77**

1 Introduction

Exploring causal effects is at the core of empirical sciences. When one draws causal inference from data by comparing groups with different treatments, "other thinks equal" is the targeted condition, meaning that groups should not systematically differ from each others in any ways, except with respect to their treatment status. A *randomized controlled trial* (RCT) has been the golden standard for causal inference for millennia, but in many situations it is not possible to run an experiment [Pearl, 2009c]. In RCTs, researchers have control over the treatment assignment of interest. In such cases, a simple comparison of the outcomes between a treated and a control group is enough for drawing a causal inference: this result is an *average treatment effect* (ATE) [Fisher, 1925, Splawa-Neyman et al., 1923, 1990, Rubin, 1974]. In the world beyond experiments, researchers cannot control the assignments, and naive comparisons are typically far from

"other thinks equal". However, by knowing the causal structures that are affected or that affect the assignment of the treatment and by controlling them, it is still possible to draw causal inference.

Econometrics alongside other applied empirical sciences have developed methods to handle *observational* (or *nonexperimental*) data with an aim to answer causal questions. In some cases, it is possible to apply identification strategies such as instrumental variables and differences-in-differences for drawing a causal inference [e.g. Angrist et al., 1996, Angrist and Krueger, 1999, Angrist and Pischke, 2008, Wooldridge, 2015]. Still in many cases these strategies are not applicable, and one must consider which data would be enough for controlling the assignment mechanism. In theory, it is possible to know which dataset is sufficient for identifying causal effects if the underlying causal structures are known. In this thesis, this topic is covered in the *Structural Causal Models* (SCM) framework. Another perspective for causal inference along with SCM framework is the *Potential Outcomes* (PO) framework, which has taken its place as the leading theoretical base for causal inference in applied econometrics. These two frameworks can be seen as complementary, but in some cases these frameworks give practical suggestions that are contradictory.

The fast development of causal inference is not limited only to the use of observational data. Another area where progress has been significant during the second decade of the 21^{st} century is in heterogeneous treatment effect estimation. Here the aim is not only to find the ATE in the target population but to estimate a personal treatment effect for an individual with some observed features. In this area, machine-learning methods have shown their applicability. Even if off-the-shelf algorithms cannot be directly applied to the estimation tasks, the modified versions have performed well not only with experimental data but also with observational data.

The prior covariate selection for the heterogeneous treatment effect algorithms has not been in focus in earlier literature. Typically the assumed condition for treatment effect estimation is unconfoundedness, meaning that the assignment of the treatment is independent of the potential outcomes. The approach that has been suggested in PO literature is to balance the covariate distributions as far as it is possible, which means in practice that one should include all the observed relevant pre-treated variables to the model [Rubin, 2007, Imbens and Rubin, 2015]. However, as it is shown in SCM literature, some causal structures cause bias in the causal estimates when wrong variables are conditioned [Pearl, 1995, 2009c].

The aim in this thesis is to explore, how causal structures affect covariate selection in the heterogeneous treatment effect estimation context. The heterogeneous treatment effect context is especially interesting because the same causal structures can affect not only the assignment and outcomes but also the treatment effect. The research will be carried out in two stages:

1. By exploring the relevant causal inference literature and by comparing their views.

2. By executing a simulation study, where a heterogeneous treatment effect

estimation method is applied with different causal structures. In each simulation, different sets of conditioned covariates are tested.

The simulation study gives insight on how the estimation performance changes with respect to the given covariate sets and underlying causal structures. Based on reviews on existing literature, it seems that no papers that consider this topic exist. This thesis provides practical recommendations for covariate selection in heterogeneous treatment effect estimation based on the theoretical perspective and the simulation study results.

2 Review of Literature

The review of literature in this thesis has the following aims:

1. To familiarize a reader with treatment effects and the fundamentals of treatment effect estimation in subsections 2.2–2.8. These subsections are based on the context of the PO framework. This framework is the base for the heterogeneous treatment effect estimation methods presented in subsections 2.17–2.21. The methodological differences between treatment effect estimation in RCTs 2.4 and in observational studies 2.5 will be clearly stated. The necessary assumptions for the causal estimation will be presented, as well as some estimation methods to use when these assumptions hold.

2. To introduce the concept of SCMs in subsections 2.9–2.15. By using tools from SCM literature it is possible to define a sufficient set of adjusted covariates for causal inference if the underlying causal structure is known. In subsection 2.15, PO and SCM frameworks will be compared, especially with a focus on how these approaches come up in covariate adjusting issue. In the simulation study the SCM approach will be applied in the data generating process.

3. To give an overview for a reader on heterogeneous treatment effect estimation methods in subsection 2.17. The *causal forest* method will be covered in more detail in subsection 2.21 because it will be used in the simulation study. Also the building blocks of the causal forest algorithm will be explained in subsections 2.19 and 2.20. The subsection 2.22 provides a glance at simulation studies in heterogeneous treatment effect estimation literature.

In this thesis treatment can be defined as an action with the aim to affect the output of interest. For example in medicine, treatment can mean the use of drugs or exercises to cure a person of an illness or injury. In economics, the term treatment is borrowed from medical trials and means actions such as taking an insurance, when the treatment effect would be the effect on the health of a person [e.g. Finkelstein et al., 2011], or the treatment effect of the training on earnings [e.g. Heckman and Robb, 1985]. A treatment can be formalized to be

some categorical or numerical variable W, but this study focuses only on binary treatments $W \in \{0, 1\}$, where $W = 1$ means treated and $W = 0$ correspondingly nontreated.

2.1 The Framework of Potential Outcomes

The PO framework [Fisher, 1925, Splawa-Neyman et al., 1923, 1990, Rubin, 1974, 1978] has been one of the two mathematical languages for causality alongside the *simultaneous equations models*. Both languages have their roots in the work of Haavelmo [1943] and have been part of statistics and especially econometrics ever since. It has been shown that these two are mathematically equal [Holland, 1988, Pratt and Schlaifer, 1988, Pearl, 1995, 2009c]. The PO framework has taken its place as the primary approach in causal inference literature in economics since the 1990s and has shown its applicability [LaLonde, 1986, Angrist, 1990, Angrist and Krueger, 1991, Krueger and Ashenfelter, 1992, Card and Krueger, 1993]. The simultaneous equations modeling approach is not covered as its own topic in this thesis, but some of its features will be discussed as a part of SCM in subsections 2.9–2.15.

All the heterogeneous treatment effect methods presented in this thesis are based on the PO framework. When these methods are applied to observed data, they aim to "mimic" RCTs similarly as described in subsections 2.6 and 2.7. All these methods are assuming unconfoundedness (subsection 2.5), which is discussed both in PO and SCM frameworks perspectives.

2.2 Treatment Effect and the Framework of Potential Outcomes

The following framework to analyze potential outcomes of a treatment has its origins in the works of Fisher 1925 and Neyman [1923, 1990]. The approach presented in this subsection is based on the work of Rubin [1974, 1978] and is named *Rubin Causal Model* (or Neyman-Rubin Causal Model) by Holland [1986]. In this framework the causal effect can be represented with the following example: Assume that there is a unit (or an individual) who can be set on the binary operation $W \in 0, 1$, for example, to get treated with a drug ($W = 1$) or not ($W = 0$). The potential outcomes are $Y(1)$ when the individual is treated and $Y(0)$ when the individual is not treated:

$$Y^{Obs} \equiv Y(W) = \begin{cases} Y(0) & \text{if } W = 0 \\ Y(1) & \text{if } W = 1 \end{cases}$$

The causal effect of the treatment τ is the difference of the potential outcomes.

$$\tau \equiv Y(1) - Y(0)$$

The problem in obtaining the τ is that only one of the outcomes can be observed at once. This problem has been called the *"fundamental problem of causal*

inference" [Holland, 1986]. This means that obtaining τ for any individual is impossible. However, in certain conditions, the average treatment effect can be observed.

2.3 Average Treatment Effect

Consider N units, indexed by $i = 1, ..., N$. Each unit i is exposed to a binary treatment $W_i \in \{0, 1\}$. Let $\boldsymbol{W}$ be a vector including all N treatment indicators for the observations. With no further assumptions, the outcome for a unit i can depend on all the N individual treatments, meaning 2^Npotential outcomes $Y_i(\boldsymbol{W})$. This is an issue in many cases: consider a case of vaccinations against infectious diseases where a proportion of the treated population affects the probability of infection for every individual in the population. Still, in many cases it is reasonable to assume that the outcome for a unit i is only dependent on the treatment W_i it has received. This assumption is called *stable unit treatment value assumption*, which means that for every N units the realized outcome can be expressed as

$$Y_i^{Obs} = Y_i(W_i) \tag{2.1}$$

Expressed in this way, the treatment effect for unit i is $\tau_i = Y_i(W_i = 1) - Y_i(W_i = 0)$, where either $Y_i(W_i = 1)$ or $Y_i(W_i = 0)$ is observed and the other one is a counterfactual outcome. The observed outcome for unit i can now be rewritten as the following:

$$\begin{aligned} Y_i^{Obs} &= \begin{cases} Y_i(0) & \text{if } W = 0 \\ Y_i(1) & \text{if } W = 1 \end{cases} \\ &= Y_i(0) + [Y_i(1) - Y_i(0)]W_i \end{aligned}$$

Thus, by assuming the *strong law of large numbers* (SLLN), the difference in means $\Delta\mu$ of the treated (units i with $W_i = 1$) and control (units i with $W_i = 0$) group is the following:

$$\begin{aligned} \Delta\mu &\overset{\text{SLLN}}{=} \mathbb{E}[Y_i \mid W_i = 1] - \mathbb{E}[Y_i \mid W_i = 0] \\ &= \mathbb{E}[Y_i(1) \mid W_i = 1] - \mathbb{E}[Y_i(0) \mid W_i = 0] \end{aligned} \tag{2.2}$$

This expression is needed to show that the simple difference in means of the groups is the sum of the ATE $\mathbb{E}[\tau_i] = \mathbb{E}[Y_i(1) - Y_i(0)]$ and the *selection bias*. A counterfactual term $\mathbb{E}[Y_i(0) \mid W_i = 1]$ (or symmetrically $\mathbb{E}[Y_i(1) \mid W_i = 0]$) can be added and subtracted in equation 2.2 so that the result will not change.

$$\begin{aligned} \Delta\mu &\overset{\text{SLLN}}{=} \mathbb{E}[Y_i \mid W_i = 1] - \mathbb{E}[Y_i \mid W_i = 0] \\ &= \mathbb{E}[Y_i(1) \mid W_i = 1] - \mathbb{E}[Y_i(0) \mid W_i = 1] \\ &\quad + \mathbb{E}[Y_i(0) \mid W_i = 1] - \mathbb{E}[Y_i(0) \mid W_i = 0] \end{aligned}$$

In this expression the ATE for the treated group is

$$\begin{aligned}\mathbb{E}[\tau_t] &= \mathbb{E}[Y_i(1) - Y_i(0) \mid W_i = 1] \\ &= \mathbb{E}[Y_i(1) \mid W_i = 1] - \mathbb{E}[Y_i(0) \mid W_i = 1]\end{aligned}$$

(or the ATE for the control group $\mathbb{E}[\tau_c]$ in a similar way) and part

$$\mathbb{E}[Y_i(0) \mid W_i = 1] - \mathbb{E}[Y_i(0) \mid W_i = 0]$$

(or symmetrically $\mathbb{E}[Y_i(1)|W_i = 0] - \mathbb{E}[Y_i(1)|W_i = 1]$) can be interpreted as the selection bias.

The selection bias means that the properties of the treated and nontreated populations are systematically differing from each other by their characteristics, which makes the naive difference in means between the groups a bad estimator for the ATE. The cause of this bias is *confounding*: the assignment system is selective and not independent of the potential outputs $\{Y_i(0), Y_i(1)\}$. This is a typical problem when the causal treatment is estimated from observational data. An economic example could be the effect of attendance at a private school W on future income Y, where the compounding variable X might include variables such as parental income and cognitive capability, which affect both treatment assignment and future income. This kind of confounding would make naively estimated treatment effect τ upwardly biased [Dale and Krueger, 2002].

2.4 Randomized Controlled Trials

The selection bias must be removed to find the causal inference. The traditional approach to do this is to organize a RCT, often stated as the golden standard of causal estimation, where the assignment operation is randomized. This makes every assignment decision W_i independent of the potential outcomes $\{Y_i(0), Y_i(1)\}$:

$$Y_i(0), Y_i(1) \perp\!\!\!\perp W_i \tag{2.3}$$

Because the assumption 2.3 holds in RCTs, the difference in means $\Delta\mu$ between the treated and the control groups equals the ATE under the stable unit treatment value assumption 2.1:

$$\begin{aligned}\Delta\mu &\overset{\text{SLLN}}{=} \mathbb{E}[Y_i(1) \mid W_i = 1] - \mathbb{E}[Y_i(0) \mid W_i = 0] \\ &= \mathbb{E}[Y_i(1) \mid W_i = 1] - \mathbb{E}[Y_i(0) \mid W_i = 1] \\ &= \mathbb{E}[Y_i(1) - Y_i(0) \mid W_i = 1] \\ &= \mathbb{E}[Y_i(1) - Y_i(0)] \\ &= \mathbb{E}[\tau]\end{aligned} \tag{2.4}$$

The estimator for ATE $\mathbb{E}[\tau]$, which is proposed in Splawa-Neyman et al. [1923, 1990] is the difference in observed average outcomes between treated and

control groups:

$$\hat{\tau} = \frac{1}{N_t} \sum_{i:W_i=1} Y_i^{Obs}(1) - \frac{1}{N_c} \sum_{i:W_i=0} Y_i^{Obs}(0) \tag{2.5}$$

Properties of this estimator can be read from the appendices (A and B).

2.5 Observational Data and Unconfoundedness

In observational studies, selection bias can be completely eliminated only if the treatment assignment system is fully known and the covariates affecting it are all obtained, whereas in RCTs the randomization removes exceptionally all bias, both observed and unobserved [Rubin, 2007]. The first problem is that the assignment system is typically unknown. On the other hand, organizing a RCT is impossible in many cases, especially in social sciences. Still, in many cases when a RCT cannot be made, it is possible to draw a causal inference if the covariates breaking $W_i \perp\!\!\!\perp (Y_i(0), Y_i(1))$ can be controlled. In this way, the assumption 2.3 can be generalized by conditioning the treatment W_i with X_i. This assumption is called *unconfoundedness*, also known as *strong ignorability* and *conditional independence assumption*:

$$Y_i(0), Y_i(1) \perp\!\!\!\perp W_i \mid X_i \tag{2.6}$$

This assumption means that in a group with the same values of covariates $X = x$, the treatment assignment can be assumed to be random. To be able to estimate the treatment over the whole covariate space, the *overlapping condition* must hold:

$$0 < \mathbb{P}[W_i = 1 \mid X_i = x] < 1 \tag{2.7}$$

Similarly as in the RCTs, by assuming 2.6, it is possible to form two conditional mean functions, which are the expected response given treatment and the expected response given control:

$$\mu_1(x) \equiv \mathbb{E}[Y_i \mid X_i = x, W_i = 1]$$

and

$$\mu_0(x) \equiv \mathbb{E}[Y_i \mid X_i = x, W_i = 0] \tag{2.8}$$

By using these two conditional means $\mu_1(x)$ and $\mu_0(x)$, we can construct the *conditional average treatment effect* (CATE) $\tau(x)$, which is the main effect of interest in this thesis:

$$\tau(x) \equiv \mu_1(x) - \mu_0(x) \tag{2.9}$$

This expression can be interpreted as the answer to the following question: *"Given the covariate values $X = x$ for an individual i, what is the expected treatment effect $\tau(x)$?"*. If the output is binary, $Y_i \in \{0, 1\}$, the expectation of Y_i is simply the probability of observing $Y_i = 1$.

2.6 Methods to Remove Selection Bias from Observational Data

The unconfoundedness assumption 2.6 states the required condition for drawing a causal inference, but does not describe how the conditioning should be applied and which set of covariates should be conditioned. The list of robust approaches with respect to the requirements of the unconfoundedness assumption include methods such as *matching*, *subgrouping*, and regression approaches with identification strategies such as *instrumental variable methods*, *differences-in-means* method and *regression discontinuity designs*. All of these methods are in the econometrician's toolkit (Angrist and Pischke [2008] as a good overview on how these strategies are applied in economics). In this thesis the focus more on matching and subgrouping methods that are applied in heterogeneous treatment effect estimation methods.

It is not easy to show that the unconfoundedness assumption would hold: the unconfoundedness assumption 2.6 is not testable without further assumptions of the causal structures. The data itself does not have information that would reveal that does unconfoundedness assumption hold or not not. However, it is possible to "assess" unconfoundedness if there are multiple control groups [Rosenbaum et al., 1987] or lagged outcomes [Heckman and Hotz, 1989], but these methods are not covered in this thesis. Also, based merely on the obtained data, it is impossible to know which variables should be selected to be conditioned [Pearl, 2009c]. In section 2.9 these questions will be covered in an asymptotic context by using prior knowledge about the causal structures. Before this, some PO strategies with an aim of filling the unconfoundedness assumption are presented.

According to Imbens and Rubin [2015], a good practice in covariate selection is to balance covariate distributions: this state is called *designing*. In this state, the researcher analyzes the available data, excluding the outcomes, in order to assemble samples with improved balance in covariate distributions. This is done by conditioning the outcomes with respect to (possibly a large number of) pre-treatment covariates (meaning that they precede the treatment, or that they are not themselves affected by the treatment): *"Given this set of proper pre-treatment variables, one generally wants to control for as many as possible, or all of them."* [Imbens and Rubin, 2015]. However, the practice of controlling the maximum number of pre-treatment covariates has been conceptually criticized [Pearl, 2009a, Sjölander, 2009]. The critique is covered in subsection 2.16.

There is a practical issue in using high dimensional covariate space for conditioning: drawing causal inferences becomes more challenging when the number of the conditioned covariate space dimensions increases but the number of observations is finite. Consider the case where there are two binary covariates ($\mathbb{X}^2$) and compare it to the case where there are twelve binary covariates ($\mathbb{X}^{12}$): In the case of $\mathbb{X}^2$ there are four covariate value combinations, but in the latter case the number of covariate value combinations increases to 4096, decreasing the observations per subclass with covariate values $X = x$, possibly leaving some subclasses empty or without comparable (and especially big enough) ob-

servation groups for causal inference. In this situation it is credible to find lower-dimensional functions of the covariates that are sufficient for removing the selection bias. *Balancing score functions* solve this problem. The *propensity score* is the most well known of the balancing score functions. These are covered in the following subsection 2.7.

2.7 Balancing Scores and the Propensity Score

By assuming unconfoundedness 2.6 the selection bias can be removed from comparisons between treated and control units by adjusting for differences in observed covariates X_i. Balancing score is a function of X_i, $b(X_i)$, so that the probability (in a infinite *"super-population"*, see Imbens and Rubin [2015, p. 20, 39 and 266]) of receiving the active treatment is independent of the covariates conditioned on the balancing score:

$$W_i \perp\!\!\!\perp X_i \mid b(X_i) \tag{2.10}$$

By giving the balancing score and by assuming that the 2.1, 2.6 and 2.7 hold, then the assignment is unconfounded given any balancing score (shown in Appendix C):

$$W_i \perp\!\!\!\perp Y_i(0), Y_i(1) \mid b(X_i) \tag{2.11}$$

Balancing score functions are not unique, and as an example, the simplest balancing function is the covariate vector X_i itself. The balancing scores of interest are those with low dimensions: the propensity score being the best known one. The propensity score is the conditional probability of receiving the treatment given $X_i = x$:

$$e(x) \equiv \mathbb{P}(W_i = 1 \mid X_i = x)$$

Because the propensity score is a balancing score function (shown in Appendix D), the assignment W_i is unconfounded given the propensity score $e(X_i)$:

$$W_i \perp\!\!\!\perp Y_i(0), Y_i(1) \mid e(X_i) \tag{2.12}$$

The intuition in the balancing scores is similar with the RCTs: if 2.10 holds, differences in covariate values between treated and control units do not lead to bias because they get canceled out when averaging over all units with the same value for the balancing score $b(x)$. Even though the covariate values may differ between the individuals assigned to the treated and control group with the same value for the balancing score $b(x)$, they have the same distribution of covariate values. Of course, this holds also for the propensity score: *similar* individuals (or groups of individuals) in the treated and control groups are being compared, when the similarity means that they are getting treated with approximately the same probability $e(x)$. [Imbens and Rubin, 2015]

There are multiple ways to apply the propensity score for achieving the conditional independence between treatment assignment and potential outputs

2.12. The four most applied methods are *covariate adjustment* using the propensity score, *stratification* or *subclassification* on the propensity score, *matching* on the propensity score and *Inverse Probability of Treatment Weighting* (IPTW) [Rosenbaum and Rubin, 1983a, Rosenbaum, 1987]. There are fundamental differences between these methods, although all four general approaches aim to estimate the same treatment effects: the focus of the methods vary with respect to the unknown components of the joint distribution of the potential outcomes, assignment process, and covariates [Imbens and Rubin, 2015].

Covariate adjustment using propensity score, also called model-based imputation, relies on building a model for missing potential outcomes. In practice, this is normally done by regressing the outcomes with respect to the treatment-status indicator as the independent variable, and the propensity score as the control variable. When the outcome Y is continuous, a linear model would be applied, and when the Y is dichotomous, a logistic regression would be selected. The coefficient of the treatment indicator is an estimate of the treatment effect: for a linear model, the treatment effect is an adjusted difference in means, whereas for a logistic model it is an adjusted odds ratio [Austin, 2011].

In stratification on the propensity score the observations are ordered based on their propensity scores and then divided in J subclasses. In each J subclass, the observations have approximately the same propensity scores and thus similar covariate distributions, if the propensity score is modeled correctly (one approach for estimating the propensity score is presented in Appendix E). In this case, group-specific differences in means gives a relatively good estimate of the treatment effect in each group. The weighted average over the J groups gives an estimate $\hat{\tau}^{\text{Strat}}$ for the ATE in the population.

The class of matching methods is a widely studied and applied subject in statistics [Rubin, 2006, Imbens and Rubin, 2015, Rosenbaum et al., 2010], econometrics [Imbens, 2004, Wooldridge and Imbens, 2008] and health sciences [Austin, 2008]. Originally this method has been presented in Rosenbaum and Rubin [1983a]. The matching method involves pairing and comparing treated and control observations that have "similar" covariate distributions. Matching on the propensity score is only one way of performing matching with respect to covariates X, but as in stratification , the propensity score is a relatively good measure of the "similarity". The matching process can be executed in many ways, but these methods are not covered in this thesis [see e.g. Rosenbaum, 1989, Dehejia and Wahba, 2002, Austin, 2008, Austin and Stuart, 2015, Imbens and Rubin, 2015].

The last propensity score method that is presented in here is the IPTW. In the IPTW, the propensity score estimate is applied to create an artificial population from the observations. As in the other three propensity score estimation methods, the aim is to balance covariates X [Joffe et al., 2004].The IPTW is based on the following equality:

$$\mathbb{E}[\tau] = \mathbb{E}\left[\frac{Y_i^{Obs} \cdot W_i}{e(X_i)} - \frac{Y_i^{Obs} \cdot (1 - W_i)}{1 - e(X_i)}\right]$$

The IPTW method has become a popular method among the observational

studies. When using the IPTW, one has to consider that the estimates become sensitive to minor changes in the specification of the model for the propensity score when there exists a substantial difference in the covariate distributions by treatment status. [Austin and Stuart, 2015]

2.8 Limitations of the Potential Outcome Approach

As mentioned in section 2.6, the data itself does not give information that is it confounded and from where the possible confounding between the treatment and output comes [Imbens and Rubin, 2015]. In his book Pearl [2009c] argues that by limiting ones terminology to statistical language without causal assumptions, it is not possible to formally define confoundedness. In the same book, Pearl shows that attempts to define confoundedness in statistical terms without causal assumptions have all failed: definitions for unconfoundedness proposed earlier in literature fail at either *sufficiency criterion* (criterion never errs when it classifies a case as nonconfounding) or *necessity criterion* (criterion never errs when it classifies a case as confounding), or both[1]. Without a formal definition of what confounding means, it is impossible to state whether the unconfoundedness assumption is valid for the sample.

As Pearl expresses in his book [2009c], concepts such as randomization, instrumental variable, exogeneity and intervention are similar to confoundedness in the way that they only exist in a causal context and thus are not statistical concepts. All these concepts require some causal assumptions that are not testable from observational data. As an example, if one has a joint probability distribution $f(x, y)$, one cannot test in any way whether y is originally randomized or not. Similarly, conditional probability $\mathbb{P}(\text{disease} \mid \text{symptoms})$ may reveal that a disease and some symptoms are dependent, but based on this expression it is not possible to state that "symptoms do not cause the disease". To get a deeper understanding of unconfoundedness, the SCM approach is presented in the following subsections 2.9–2.15. The SCM framework gives a way to represent causal relations between variables with structural equations and corresponding acyclic graphs. With this approach, it is possible to define explicitly the state in which unconfoundedness condition holds with a given causal graph.

2.9 Structural Causal Model

SCMs have their roots in structural equation models, *path analysis* [Wright, 1928] and *Bayesian networks* [Pearl, 1985] and are mostly associated with work by Judea Pearl and his collaborators. The SCM framework is complementary to the PO framework and these two are mathematically equivalent (in recursive systems where the solution for the g^{th} endogenous variable involves only the first g equations) [Pearl, 2009c]. Although they are mathematically equivalent, they have their own strengths that make them appropriate to different questions.

[1] In econometrics the corresponding notation is *exogeneity*, examples of association criteria in economics: Engle et al. 1983, Leamer 1985, Aldrich et al. 1993

SCM framework is a formal way to represent causal structures behind relevant features in a data set. These causal assumptions are based on prior knowledge of the subject of interest. An SCM is a nonparametric model that consists of exogenous variables $\mathbb{U}$, endogenous variables $\mathbb{V}$, and a set of functions $\mathbb{F}$ that assign a joint distribution for every variable in $\mathbb{V}$ based on the other variables in $\mathbb{V}$ and $\mathbb{U}$. In an SCM, X is defined to be a direct cause of Y if it appears in the function of Y, f_Y. Correspondingly, X is a cause of Y if it is a direct cause of Y, or of any cause of Y. For every SCM, there exists a *graphical causal model* G, that contains a node for every M variable in $\mathbb{V}$ and $\mathbb{U}$. [Pearl, 1995]

2.10 Basic Terminology for Graphical Causal Models

If X is a direct cause of Y, graph G has a *directed edge* (an arrow) between the nodes representing X and Y in the following way:

$$X \longrightarrow Y \tag{2.13}$$

In 2.13 X and Y are *adjacent*, which means that they are connected with an edge: X is a *parent* of Y and correspondingly Y is a *child* of X.

When two nodes are connected with an unbroken, nonintersecting sequence of directed edges, so that the route is possible to be traced along or against the arrows, they are connected with a *path*. If two nodes are connected with a *directed path*, meaning that a route between these nodes can be traced along the arrows, the nodes (X and Y in 2.14) before some other node (Z in 2.14) are the *ancestors* of that node, and correspondingly all the nodes (Y and Z in 2.14) that come after some other node (X in 2.14) on the path are the *descendants* of that node:

$$X \longrightarrow Y \longrightarrow Z \tag{2.14}$$

In this directed path, X is a cause of Y and Z. In general, a directed path with three nodes with one directed edge into, and one directed edge out of, the middle node, as in 2.14, is called a *chain*. The other two possible ways to construct causal paths between three nodes are the *fork*:

$$X \longleftarrow Y \longrightarrow Z \tag{2.15}$$

and the *collider*:

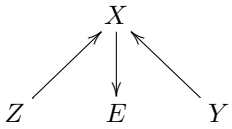
$$X \longrightarrow Y \longleftarrow Z \tag{2.16}$$

In the cases of the chain (2.14) and the fork (2.15) nodes X and Z are dependent if Y is not conditioned, whereas they are conditionally independent when node Y is conditioned ($X \perp\!\!\!\perp Z \mid Y$). In the case of the collider (2.16) the logic goes the other way around: By not conditioning Y, X and Y are independent, but by conditioning Y, X and Y are become conditionally dependent ($X \not\perp\!\!\!\perp Z \mid Y$). As mentioned earlier, in the chain structure X is a cause of both Y and Z, so it is intuitive that the three variables are likely[2] dependent. The

[2] *"Likely"* in the implications means the *intransitive case* which is the only exception for the implication. This case is rare and can be ignored in practice [Pearl, 2009c].

intuition behind the logic why conditioning the middle node (Y in 2.14) in a chain causes the conditional independence of the other two nodes (X and Z in 2.14) is that the probability distributions of X and Z are compared in each situation $Y = y$ separately, making their distributions independent. Similarly in the case of forks, the end nodes X and Z have a *common cause* Y making their probability distributions likely dependent, even if they were conditionally independent. When the common cause Y is conditioned, X's and Z's probability distributions are compared in each situation $Y = y$ separately, as in the case of the chain.

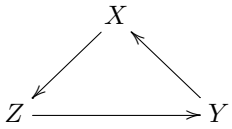
The collider's unconditional independence of the path's end nodes is intuitive: a change in X (Z) does not have an effect on Z (X), but has an effect on Y. In contrast, it may not be as clear why conditioning a collider Y makes two independent variables (here X and Z) conditionally dependent. The simplest example is the following: consider that the middle node Y is a linear combination of X and Z, $f_Y(X, Z, U_Y) = \beta_X X + \beta_Z Z$. By conditioning the Y to y, one "forces" the other variable to compensate a change in the other variable. For example, if $\beta_X = \beta_Z = 1$ and $Y = 7$, it is clear that $Z = 4$ if $X = 3$. Just as conditioning a collider, also conditioning a descendant of a collider (E in 2.17) makes previously independent variables (X and Z in 2.17) to be (likely) conditionally dependent ($Z \not\perp\!\!\!\perp Y \mid E$ in 2.17):



(2.17)

2.11 Directed Acyclic Graphs

Directed Acyclic Graphs (DAG) are the primary interest when using the SCM framework for drawing causal inference. The term acyclic means that the graph G does not have any paths from any node to itself, as there is in 2.18 as an example ($X \to Z \to Y \to X$ or similarly for Z and Y):



(2.18)

It is rather clear for economists that acyclicity excludes some systems from the scope of the DAG analysis, such as equilibrium mechanisms, which are in the core of market mechanisms. In contrast, the PO framework provides a natural language for simultaneous mechanisms such as competitive market equilibrium [Imbens, 2019].

DAG is an economic way to represent conditional independence assumptions and qualitative causal influences. Since all the independent unobserved variables can be lumped together into a set $\mathbb{U}$ characterized with a joint distribution $\mathbb{P}(u)$,

the full specification of an SCM consisting of variables $\mathbb{X}^n = \{x_1, \ldots, x_n\}$, will have two components: a set of functional relationships

$$x_i = f_i(pa_i, u_i), i \in 1, \ldots, n \qquad , \text{ where } pa_i \tag{2.19}$$

represents the parents of variable x_i, and a joint distribution $\mathbb{P}(u)$ on the background factors. A useful feature with the DAGs is the *product decomposition rule* for the joint distribution of variables included in SCM :

$$\mathbb{P}(x_1, \ldots, x_n) = \prod_i \mathbb{P}(x_i \mid pa_i), i \in 1, \ldots, n \tag{2.20}$$

The joint distribution 2.20 characterizes the graph G and allows one to derive a *post intervention distribution* of an SCM, which will be presented in 2.13. In the following subsection 2.12, the reader will become familiar with the *d-separation criterion*, which is a graphical approach to determine conditional relations of variables for any given DAG.

2.12 The d-Separation Criterion

For testing independence between nodes X and Y in the given graph G, one can use the d-separation criterion. This happens by considering all the paths between X and Y and by clarifying, whether these paths are d-separated, in other words, whether the flows of dependencies are "blocked" with a conditioned set of notes $\mathbb{Z}$. The following definition for d-separation criterion uses the basic path structures presented in 2.10 In Pearl [2009c], the definition of the criterion is given as the following (a direct citation):

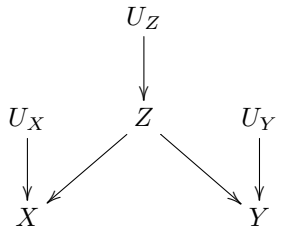
Definition. ***(d-Separation)*** *A path p is said to be d-separated (or blocked) by a set of nodes $\mathbb{Z}$ if and only if*

1. *p contains a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$ such that the middle node m is in $\mathbb{Z}$, or*

2. *p contains a collider $i \to m \leftarrow j$ such that the middle node m is not in $\mathbb{Z}$ and such that no descendants of m is in $\mathbb{Z}$.*

A set $\mathbb{Z}$ is said to d-separate X from Y if and only if $\mathbb{Z}$ blocks every path from a node in X to a node in Y.

Consider a canonical example between sales of ice cream X and drownings Y. These two variables are *d-connected* by a "back-door path" through the variable Z, temperature:



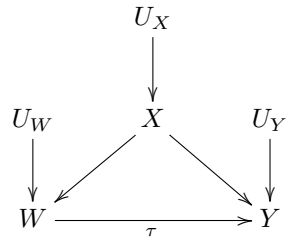
(2.21)

By conditioning the node Z, nodes X and Z are d-separated and thus conditionally independent.

The d-separation condition allows one to test independence implications in a model G. This is called *local testing*. Local testing is a nonparametric approach, in which one tests, does the data support the independence assumptions that are stated on the given DAG. As an example, if one assumes that the DAG 2.21 represents the real world, the $X \perp\!\!\!\perp Y \mid Z$ should hold. If this is not the case, one can assume that the DAG 2.21 is not correct.

2.13 Interventions in DAGs

Consider a DAG consisting of a single observed confounder X that affects the treatment assignment W and the output Y. The DAG has also exogenous variables U_W, U_X and U_Y:



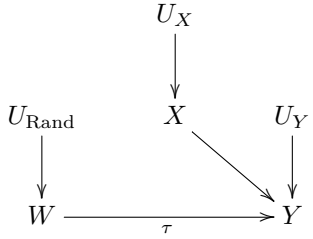
(2.22)

The corresponding SCM would be the following :

$$\mathbb{V} = \{X, W, Y\}, \mathbb{U} = \{U_X, U_W, U_Y\}$$

$$\mathbb{F} = \begin{cases} f_X(U_X) \\ f_W(X, U_W) \\ f_Y(X, W, U_Y) \end{cases}$$

There exists a "backdoor path" from W to Y, $W \leftarrow X \rightarrow Y$ making the treatment effect confounded (selection bias). All the confounding would be eliminated by organizing a RCT, as it is presented in 2.4. With a DAG 2.23 where the randomizing operation for assignment is represented with U_{Rand}, it can be directly seen that no "back-door path" exists from W to Y, thus the assumption 2.3 holds:



(2.23)

$$\mathbb{F}' = \begin{cases} f_X(U_X) \\ f_W(U_{\text{Rand}}) \\ f_Y(X, W, U_Y) \end{cases}$$

When an RCT cannot be organized, different identification strategies must be applied. As stated in the unconfoundedness assumption (2.6), with correct adjustments, it is still possible to draw a causal inference. In the context of SCMs, this is done by simulating an intervention, with *do-calculus.*

The *intervention* (or treatment) in DAG literature is represented with the notation $do(X_i = x_i)$, or $do(x_i)$ in short, meaning that variable X_i is "forced" to take a value x_i. By using this notation, the ATE is expressed as $\mathbb{E}(Y \mid do(W = 1)) - \mathbb{E}(Y \mid do(W = 0))$, where Y is the output variable and W is the treatment variable. It is valuable to consider the differences between the conditional probability $\mathbb{P}(Y = y \mid X_i = x_i)$ and the post intervention probability $\mathbb{P}(Y = y \mid do(X_i = x_i))$. In the former of these two expressions, one is observing the probability of Y to get the value y in the population where the variable X_i has a value x_i, whereas in the latter, one is obtaining the probability for $Y = y$ when the variable X_i is fixed to x_i for everyone in the population.

The idea in do-calculus is to modify the existing model $\mathbb{P}(x_1, \ldots, x_n)$ into a modified version $\mathbb{P}_m(x_1, \ldots, x_n)$ by performing a *surgery* on the graph, so that the causal effect of X on Y can be derived from the modified model by using standard probabilistic operations. The three do-calculus rules are presented in Appendix F (proofs and explanations are provided in Pearl [1995]). In other words, when it is possible to draw a causal inference (this can be clarified by using do-calculus rules), one can find a modified model of the original DAG in which the distribution $\mathbb{P}_m(Y \mid X)$ is the same as $\mathbb{P}(Y \mid do(X))$ by using a combination of the assumptions embodied in the given DAG and probabilistic tools such as conditioning. In the concept of the SCMs, the causal effect is defined in the following way [Pearl, 2009c] (a direct citation):

Definition. (Causal Effect) *Given two disjoint sets of variables, X and Y, the causal effect of X on Y, denoted as $\mathbb{P}(y \mid do(x))$, is a function from X to the space of probability distributions on Y. For each realization x of X, $\mathbb{P}(y \mid do(x))$ gives the probability of $Y = y$ induced by deleting from the model of 2.19 all equations corresponding to variables in X and substituting $X = x$ in the remaining equations.*

Graphically, this definition of a causal effect corresponds to a subgraph of the original graph G where all the directed edges pointing to X have been removed. This leads to a transformation of 2.20 which is called a *truncated factorization*

formula:

$$\mathbb{P}(x_1, \ldots, x_n \mid do(x_i')) = \begin{cases} \prod_{j \neq i} \mathbb{P}(x_j \mid pa_j) & \text{if } x_i = x_i' \\ 0 & \text{if } x_i \neq x_i' \end{cases}$$

$$= \begin{cases} \frac{\mathbb{P}(x_1, \ldots, x_n)}{\mathbb{P}(x_i' \mid pa_i)} & \text{if } x_i = x_i' \\ 0 & \text{if } x_i \neq x_i' \end{cases} \tag{2.24}$$

$$= \begin{cases} \mathbb{P}(x_1, \ldots, x_n \mid x_i', pa_i) \cdot \mathbb{P}(pa_i) & \text{if } x_i = x_i' \\ 0 & \text{if } x_i \neq x_i' \end{cases} \tag{2.25}$$

In 2.24 the joint distribution of the model is divided by a propensity score $\mathbb{P}(X_i = x_i' \mid PA_i = pa_i)$. From 2.25 one can derive an *adjustment formula for direct causes*:

$$\mathbb{P}(y \mid do(x_i')) = \sum_{pa_i} \mathbb{P}(y \mid x_i', pa_i) \cdot \mathbb{P}(pa_i) \qquad \text{, where } Y \tag{2.26}$$

is any set of variables disjoint of $\{X_i \cup PA_i\}$.

For using 2.26 to draw causal inference, all the parents of X should be observed. Usually this is not the case. By applying the three do-calculus rules, it is possible to derive the two most important causal effect identification strategies for DAGs: the *back-door criterion* and the *front-door criterion*. In this thesis, only the former criterion is covered because it can be applied to define if a set of conditioned variables $\mathbb{Z} \subseteq \mathbb{V}$ is sufficient for identifying $\mathbb{P}(y \mid do(x))$ [Pearl, 1995].

2.14 The Back-Door Criterion

As stated earlier in subsection 2.8, the PO framework does not provide any definition when the obtained covariate set $\mathbb{Z}$ is sufficient to fill the unconfoundedness assumption 2.6. In the SCM framework, it is possible to test whether conditioning a covariate set $\mathbb{Z}$ is sufficient for identifying $\mathbb{P}(y \mid do(x))$ for a given DAG G with the back-door criterion. In Pearl [1995], the back-door criterion for a given DAG G is defined in the following way (direct citation):

Definition. (Back-Door) *A set of variables $\mathbb{Z}$ satisfies the back-door criterion relative to an ordered pair of variables (X_i, X_j) in a DAG G if:*

(i) no node in $\mathbb{Z}$ is a descendant of X_i; and

(ii) $\mathbb{Z}$ blocks every path between X_i and X_j that contains a directed edge into X_i.

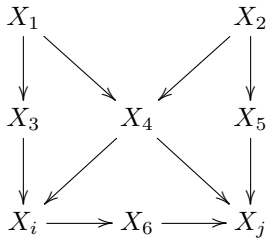
Similarly, if X and Y are two disjoint subsets of nodes in G, then $\mathbb{Z}$ is said to satisfy the back-door criterion relative to (X, Y) if it satisfies the criterion relative to any pair (X_i, X_j) such that $X_i \in X$ and $X_j \in Y$.

With a set $\mathbb{Z}$ filling the back-door criterion, the causal (or treatment) effect of X on Y is given in the following way:

$$\mathbb{P}(y \mid do(x)) = \sum_{z} \mathbb{P}(y \mid x, z) \cdot \mathbb{P}(z) \tag{2.27}$$

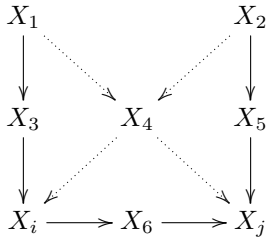
The proof for the *back-door adjustment formula* is provided in Pearl [1993].

As an example, in DAG 2.22 the sufficient set $\mathbb{Z}$ for estimating the ATE from W to Y is $\{X\}$ since X is not a descendant of W and it blocks the only path with a directed edge into W between W and Y, $\{Y \leftarrow X \rightarrow W\}$. In a more complex graph 2.28, some sufficient sets in $\mathbb{Z}$ for estimating $\mathbb{P}(x_j \mid do(x_i))$ are $\{X_3, X_4\}$ and $\{X_4, X_5\}$ (the DAG is provided in Pearl [2009c]. Exogenous variables $\mathbb{U}$ are not drawn into the DAG to keep it more accessible). In contrast, adjusting for $\{X_4\}$ or $\{X_6\}$ would lead to a biased estimate:



(2.28)

If the variable covariate X_4 were not observed there would not be any sufficient set $\mathbb{Z}$ to fill the back-door criterion because the back-door path $\{X_j \leftarrow X_4 \rightarrow X_i\}$ cannot be blocked[3]:



(2.29)

The back-door adjustment formula (2.27) typically leads to similar results as adjusting by matching (with or without propensity score), weighting (IPTW), adjusting with propensity score or by regression, and by stratifying (with or without propensity scores) [Imbens, 2019]. The difference in these is that the covariate selection can be done systematically, if the assumed DAG represents approximately the real world.

2.15 Limitations of the SCM Approach

In his essay, Imbens [2019] gives a critical overview of the SCM approach with an economic perspective. In his essay the main critique concerns the hardship to find real independencies, especially in social sciences. As stated in Gelman and Imbens [2013]: *"More generally, anything that plausibly could have an effect will not have an effect that is exactly zero."*. In many cases DAGs give

[3]In contrast, $\mathbb{P}(x_j \mid do(x_i))$ in DAG 2.29 can be solved by using the *front-door criterion*, which is presented in Pearl [1995]

testable independence implications, however in general this is not the case. The main advantage of using formal graphical methods to identify causal effects is achieved when models become more complex. However, in modern economics, researchers typically avoid using too complex causal models with a big number of variables, and lean more towards credible identification strategies, such as natural randomized experiments [Angrist and Pischke, 2010]. An article *"Let's Take the Con out of Econometrics"* by Leamer [1983] has significantly influenced this tendency.

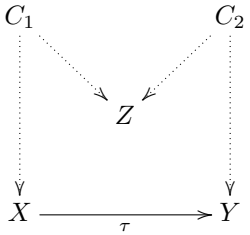
Imbens [2019] lists other possible reasons why the SCM framework has not assumed a bigger role in econometrics. One is shape restrictions, such as monotonicity, convexity or concavity, which can be integrated more easily with the PO framework than with the SCM approach [Angrist et al., 1996, Matzkin, 1991, Chetverikov et al., 2018]. The second mentioned reason is that the PO framework is connected very naturally to many traditional economic topics such as demand and supply, where potential outcomes are theoretical primitives. The third reason is that the PO framework has been successful in heterogeneous treatment effect estimation, where in contrast the SCM framework has yet not shown its suitability. In addition, when comparing SCMs to SEMs, only the latter can be used in simultaneous cases, as stated in subsection 2.11 [Heckman, 2008].

Finally, perhaps the most convincing reason is that the PO and the SEM frameworks have been sufficient analytical tools for decades in social sciences, whereas at the same time there is a lack of empirical papers using SCMs, at least in economics. As Imbens [2019] states, *"History suggests that those are what is driving the adoption of new methodologies in economics and other social sciences, not the mathematical elegance or rhetoric."*. Time will show whether the economic community will find the SCM framework to be useful on its own or when it is partially integrated with other methods [e.g. Moneta, 2008]: all the empirical methods that are nowadays part of an econometrician's toolkit have had to have a critical mass of empirical evidence of their applicability, and this might take time. In this thesis, the SCM approach will be used as the basis in the data-creating process in the simulation study, which allows one to formally define the sufficient conditioned set of variables for unconfoundedness.

2.16 When to Adjust?

In the simulation study 2.23 the back-door criterion 2.14 is used to define whether the set of adjusted variables $\mathbb{Z}$ is sufficient to make causal inference unconfounded. This can be done directly because the underlying SCM is known. But as stated in subsection 2.15, finding the underlying (approximate) SCM is far from easy in the real world. In the PO literature, the suggestion has been to balance covariate distributions as far as one can by conditioning *"as many as possible"* proper pretreated variables [Imbens and Rubin, 2015] (more in subsection 2.6). This approach has been criticized in SCM literature: a key example has been a so-called *M-bias* [Sjölander, 2009, Pearl, 2009a]. The DAG for the M-structure is the following (exogenous variables $\mathbb{U}$ are left out from the DAG

for simplicity):



(2.30)

Here one would like to estimate the effect from X to Y. In addition to these two variables, variable Z is also observed whereas variables U_1 and U_2 are left unobserved. Variable Z is a collider and d-separates the back-door path from X to Y and thus should not be conditioned, even if it is a proper pre-treated variable.

There has been discussion in literature as to how usual the M-structure is in practice [e.g. Glymour, 2006, Kelcey and Carlisle, 2011, Liu et al., 2012, Ding and Miratrix, 2015]. This question was at the very center in a well known controversy that took place in 2007–2009, starting with an article written by Rubin [2007], which suggested balancing covariate distributions of treated and controlled groups as far as possible. This was questioned by Shrier [2008], Sjölander [2009] and Pearl [2009a] with the M-structure case (to be precise, in these papers the questioned part was the use of a propensity score including a collider in the M-structure: in this case the propensity score would be a descendant of the collider).

In his response, Rubin [2009] argued *"I cannot think of a credible real-life situation where I would intentionally allow substantially different observed distributions of a true covariate in the treatment and control groups"*, referring to the M-structure. As a response, Pearl [2009b] claimed that cases containing a local M-structure are abound: *"Every time we condition on a variable that is not causally related to both treatment and outcome but merely associated with the two, we may introduce an M-bias."*. As an example of this kind of variable, Pearl introduced seat-belt usage (Z in 2.30), which was a variable included in a study case presented originally in Rubin [2007] where the aim was to estimate the effect of smoking (X in 2.30) in relation to lung cancer (Y in 2.30). The potential latent unobserved variables could be *attitudes affecting to smoking habits* (C_1) and *attitudes affecting to susceptibility to lung diseases* (C_2):

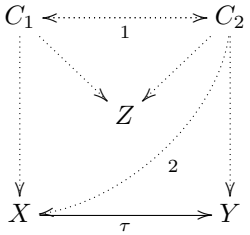
"Obviously, seat-belt usage has no causal effect on smoking or lung diseases; it is merely an indicator of a person's attitudes toward societal norms as well as safety and health related measures. Some of these attitudes may affect smoking habits, and some may affect susceptibility to lung diseases. If we have good reasons to believe that these two types of attitudes are marginally independent, we have a pure M-structure on our hand.".

This controversy has been noted in other articles. To point out how intuitively implausible the M-structure would be in this particular example, Imbens [2019] mentioned that if the DAG 2.30 represented the underlying system, it would mean that one could estimate the causal effect of smoking in relation to

lung cancer, but at the same time it would be impossible to estimate the same effect for the group of seat belt users and non-users separately. In the article written by Ding and Miratrix [2015], two possible deviations from the exact M-structure were introduced: dependent latent variables C_1 and C_2 (the directed edge 1 in DAG 2.31) and a directed edge from C_2 to X (2 in DAG 2.31). The former one would mean that the attitudes towards smoking and susceptibility to lung diseases would be dependent. In his paper, Pearl [2009b] saw that the assumption of the independence of these two variables ($C_1 \perp\!\!\!\perp C_2$) would be strong:

"But even if marginal independence does not hold precisely, conditioning on "seat-belt usage" is likely to introduce spurious associations, hence bias, and should be approached with caution."

However, as Ding and Miratrix [2015] showed by using linear SCM methods, if $C_1 \perp\!\!\!\perp C_2$ does not hold it is almost always better to condition the collider (Z in 2.31) in terms of the overall bias in the estimate for τ, excluding some extreme cases with a high correlation between the collider and the latent variables simultaneously with a mild correlation between the latent variables. Similarly, it seems implausible that the assumption $C_2 \perp\!\!\!\perp X$ would hold, which would imply erasement of the directed edge 2 in 2.31. Ding and Miratrix [2015] showed that when all the correlations between covariates in DAG 2.31 are set as equal, conditioning the Z strictly dominates the opposite strategy in terms of estimation bias. These two results are in line with previous qualitative [Greenland, 2003] and simulation [Liu et al., 2012] studies in the way that collider bias generally tends to be small in real world scenarios.



(2.31)

The study results provided above suggest the PO approach to balance covariate distributions by conditioning as many proper pre-treated variables as possible, perhaps considering the statistical efficiency, as a good practice. Another practice that is suggested in many sources in PO literature is to do *sensitivity analyses* for the estimated results [e.g. Rosenbaum and Rubin, 1983b, Imbens, 2003, Andrews et al., 2017, Andrews and Oster, 2019]. The idea in sensitivity analyses is to test how sensitive the results are for an unobserved confounder that affects both the treatment and the outcome. This topic is not covered further in this thesis.

Of the studies mentioned above, only [Liu et al., 2012] was executed as an empirical study, when all the other studies only considered asymptotic properties. Alongside asymptotic properties, also finite sample properties are highly relevant in practice. Although theoretical discussion of the finite samples is

more difficult, it is possible to get an insight into these questions with simulation studies. Subsection 2.22 gives an overview of simulation study designs used in heterogeneous treatment effect estimation literature.

2.17 Heterogeneous Treatment Effects

ATE, as a population average result has its own advantages and disadvantages: if the variability of individual causal effects is wide in a target population, an average result may be relatively uninformative when it is applied to an individual of interest. Even if in some cases the assumption of the *constant treatment effect*, also called *additivity assumption*, might be sufficient, it cannot be generally assumed [Holland, 1986]. One step beyond the average treatment point estimator is the *quantile treatment estimator*, which allows to analyze treatment effect distribution across a target population [e.g. Abadie et al., 1999, Chernozhukov and Hansen, 2005, Firpo, 2007, Bitler et al., 2006], but does not answer what the expected treatment effect is for an individual with the characteristics $x \in \mathbb{X}$. Quantile treatment estimators will not be covered in this thesis.

By understanding heterogeneity of treatment effects in a target population, relevant questions such as "which individuals benefit from the treatment?" and "how are the treatment effects affected by the covariates?" can be better answered. Also by discovering heterogeneity of treatment effects, it is in principle possible to apply the results for populations that differ by their characteristics from the originally investigated population, in other words, the results have *external validity* [Hotz et al., 2005, Athey and Imbens, 2016a]. This means that if the covariate vector $x_i \in \mathbb{X}$, which fills the unconfoundedness assumption, is obtained for a given individual i and the conditional average treatment estimate (2.9) is known, it is unproblematic to estimate the treatment effect $\tau(x_i)$ for this individual.

A traditional approach to test and handle treatment effect heterogeneity is to estimate ATEs for specified subpopulations based on a substantive interest [Athey and Imbens, 2016a]. In this approach, the manner of selecting the subgroups is not driven by data but is, instead, based on earlier knowledge about the subject. One raised problem in this approach is that subpopulations may be selected ex post, implying multiple testing concerns. Generally used approaches to avoid these concerns prevent statistical "cherry picking" by using per-analysis plans [Casey et al., 2012] or multiple testing corrections [e.g. List et al., 2016].

2.18 An Overview of Treatment Effect Heterogeneity Estimation Methods

A traditional approach in heterogeneous treatment effect estimation has been to perform subgroup analysis [Gail and Simon, 1985, Bonetti and Gelber, 2004]. One problem with this approach is to find the sufficient subgroups based on the data by avoiding multiple testing issues and statistical "cherry-picking". Additionally, typically the treatment effect is not linearly dependent on the covariates, which complicates the task. An example of a simple way to explore

heterogeneity in a data-driven way would be to specify a regression of the outcome with an indicator for the treatment status. A suggested approach for the regression would be to represent covariates as indicators and include the interaction terms between the covariate indicators and the treatment indicator in the model. This method gives an unbiased estimate for $\tau(x)$ if all covariates are given as indicators, they partition the population, and the model includes all the interactions between the indicators and the treatment indicator [Athey and Imbens, 2016a]. Using this approach is, however, only possible with low-dimensional datasets.

Typically heterogeneity is explored over a high-dimensional covariate space and thus traditional statistical methods are not generally applicable to these problems. With high-dimensional data, machine-learning methods are more applicable than traditional statistical approaches. However, due to the counterfactual nature of causality, off-the-shelf versions of algorithms cannot be directly applied in these cases. A reason why machine-learning methods are more applicable to high-dimensional data is their ability to balance accuracy in model fitting (minimizing *bias*) and to reduce excessive complexity in the model selection (minimizing *variance*) [Hastie et al., 2009, Varian, 2014, Mullainathan and Spiess, 2017, Athey and Imbens, 2019]. Too complex, namely *overfitted*, models give poor results beyond the training data. A typical approach to balance the tradeoff between bias and variance is to minimize some prediction-error function, for example the *mean squared error* (MSE), in parallel with some penalty term for model complexity, for example the number of non-zero coefficients in the model.

A simple example of a machine learning method used in heterogeneous treatment effect estimation is a version of the linear estimating approach with interaction terms mentioned above, which can be used in situations where there are potentially a large number of covariates: as far as the underlying model has (at least approximately) a high amount of observations per covariate (or an interaction term) with an important effect on the outcome (or on treatment effect heterogeneity), *regularized regressions* can be used to explore principal covariates and interaction terms with respect to outcomes and heterogeneity. From regularized regressions, the *LASSO-like* methods (*least absolute shrinkage and selection operator*) are the most widely used in exploring treatment effect heterogeneity [e.g. Imai and Ratkovic, 2013, Tian et al., 2012, Weisberg and Pontes, 2015]. To fit a LASSO-regression in obtained data, the idea is to minimize two terms at the same time with respect to a coefficient vector β: a MSE $\sum_{i=1}^{N}(Y_i-\beta^\top X_i)^2$ and a *penalty term* $\lambda\sum_{k=1}^{K}|\beta_k|$, which penalizes higher values of β. Term λ is a so-called *tuning parameter*, which gives a weight for the penalty term:

$$\min_{\beta}\left\{\sum_{i=1}^{N}(Y_i-\beta^\top X_i)^2+\lambda\sum_{k=1}^{K}|\beta_k|\right\} \tag{2.32}$$

An advantage of using LASSO-like methods is that they can lead to *sparse solutions*, meaning that coefficients of variables that do not increase prediction

accuracy are able to get the value zero. The general approach to *tune* the parameter λ is to use *out-of-sample cross-validation* [Athey and Imbens, 2019]. This means that different values of λ are tested with *k-fold cross-validation* (Appendix 4), which gives a cross-validation estimate for each λ [e.g. Hastie et al., 2009]. At the end, λ with the lowest cross-validation estimate is chosen.

Another way besides specifying a parametric model is to construct a fully nonparametric model for estimating $\tau(x)$. Nonparametric methods do not make explicit assumptions about a function form. When the idea in parametric models is to estimate parameter values by finding the best fit for the function so that the form of the function is selected prior, in nonparametric methods the whole function is estimated based on the data. A general estimating procedure of the function for nonparametric methods is based on seeking a function that minimizes the distance to the observed data points, without being too "rough" or "wiggly".

An advantage of nonparametric methods as compared to parametric methods is their flexibility. By deciding to use a parametric function that is significantly different by form as compared to the underlying function, it is not possible to get the model to fit well in the data. Due to the fact that nonparametric models do not assume anything about the form of the underlying function, they avoid the trap of wrong modeling. At the same time, nonparametric methods have some disadvantages as compared to parametric methods. Typically, nonparametric models require more data than parametric methods do: when parametric models need data for estimating (usually) a relatively small number of parameters, nonparametric models need data to accurately estimate the whole function. The second disadvantage is that due to their flexibility, nonparametric models can easily suffer from overfitting, which may decrease their reliability. [e.g. James et al., 2014]

In their article, Athey and Imbens [2016a] have listed possible goals for using non-parametric models to estimate heterogeneous treatment effects. Descriptively, by using nonparametric methods, it is possible to gain insight into which types of units have the highest and lowest treatment effects, as well as to visualize comparative statics results, all without imposing any prior restrictions. The second goal is to gain external validity for the results. As the third goal, Athey and Imbens propose personalized recommendations. Their fourth goal gives restrictions for the use of many nonparametric methods: one may want to test hypotheses and construct confidence intervals. By desiring confidence intervals, the number of potential nonparametric methods falls relatively low [Athey et al., 2018].

An example of a simple nonparametric heterogeneous treatment effect estimation method is *K-nearest neighbor matching* [Athey and Imbens, 2016a]. Here, a treatment effect estimate $\hat{\tau}(x)$ for some x is constructed by finding K nearest treated observations $\mathcal{N}_t$, where "nearness" is measured by using Euclidean distance for a covariate vector, and similarly K nearest non-treated observations $\mathcal{N}_c$. Under unconfoundedness, the treatment effect estimation $\hat{\tau}(x)$ is then constructed by averaging the output values of both treated observations $\bar{Y}(1) = \frac{1}{K}\sum_{i\in\mathcal{N}_t} Y_i$ and non-treated observations $\bar{Y}(0) = \frac{1}{K}\sum_{i\in\mathcal{N}_c} Y_i$ and then

by subtracting $\bar{Y}(0)$ from $\bar{Y}(1)$. *Kernel estimation* works in a relatively similar way with the difference that $\bar{Y}(1)$ and $\bar{Y}(0)$ are constructed by using a weighted average, which is based on the Euclidean distance for a covariate vector: the outcome values of the nearer neighbors are getting more weight than the the outcome values of the further neighbors. However, these nearest neighbor methods have performed unsatisfyingly when the dimension of covariates has been more than three [Athey and Imbens, 2016a]. This comes from the fact that all covariates are treated symmetrically. Ideally, the covariates with the highest effects on heterogeneity should have more weight. This subject will be covered in the subsection 2.21.

As the machine-learning methods mentioned above do, most of the heterogeneous treatment effect estimation methods assume that the data has come from a RCT: this list includes methods such as *FindIt* by Imai and Ratkovic [2013], which is based on *adapted support vector machines*, a method based on the *Bayesian additive regression trees* [Chipman et al., 1998] by Green and Kern [2010], another Bayesian method Taddy et al. [2014] based on the Bayesian forests and many others. However, an increasing number of new research focuses on methods applicable to observational data [Hill, 2011, Athey and Imbens, 2016b, Luedtke and Laan, 2016, Hahn et al., 2017, Nie and Wager, 2017, Powers et al., 2017, Shalit et al., 2017, Zhao et al., 2017, Wager and Athey, 2018, Athey et al., 2019]. In this thesis, the method *causal forest* [Wager and Athey, 2018, Athey et al., 2019] will be used in the simulation study 2.23 for synthetic, non-randomized data. The causal forest is chosen because it is can be applied with observational data, it provides confidence intervals for the estimates, it is computationally efficient, and the algorithm has an `R`-package in `CRAN` [Tibshirani et al., 2020]. Before going more deeply into in causal forest in subsection 2.21, its building blocks are described in the following subsections: a predicting algorithm *regression tree* in 2.19 [Breiman et al., 1984] and a modified version of the regression tree for causal estimation called *causal tree* in 2.20 [Athey and Imbens, 2016b].

2.19 Regression Tree

Tree-based methods is a nonparametric machine-learning class. This class has a wide variety of methods from which regression trees, causal trees and causal forests will be presented in this thesis. An umbrella term *classification and regression tree analysis* consists of two methods, regression tree and classification tree analysis [Breiman et al., 1984]. Alongside the classification and regression tree algorithm, there are other decision tree algorithms such as *C4.5* [Quinlan, 2014], *Chi-square automatic interaction detection* [Kass, 1980] and *multivariate adaptive regression splines* [Friedman, 1991]. In some methods, for example in the regression tree, the idea is to grow an individual tree, whereas in other methods, such as in *random forest* [Breiman, 2001] and *gradient boosting* [Friedman, 2002], one grows multiple trees and aggregates the results from individual trees.

The regression tree analyses are used in situations where the output of interest is continuous, whereas in contrast the classification tree analysis is used for

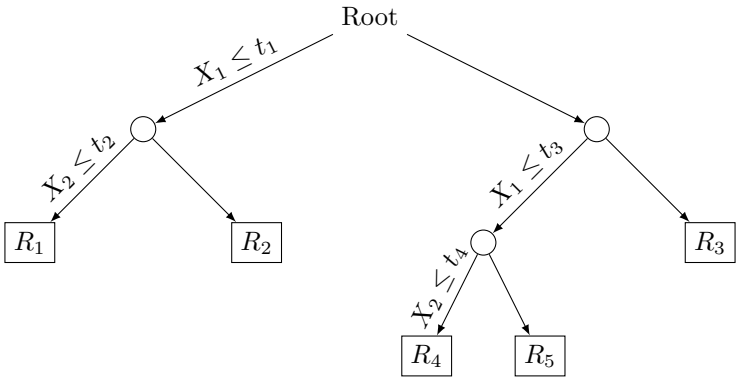


Figure 2.1: A two dimensional regression tree.

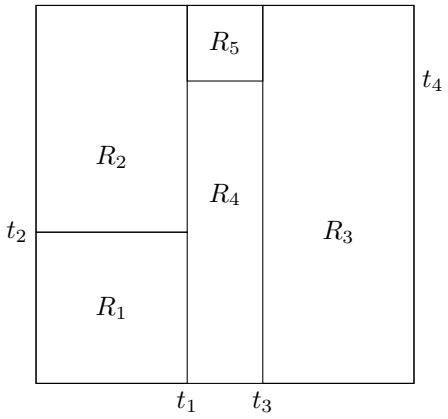


Figure 2.2: The corresponding partition of the covariate space X^2 for the regression tree 2.1.

discrete outputs. In this thesis, the focus is on continuous outputs (this is done for simplicity: all the methods that are considered in this paper are applicable also to discrete cases) and thus only the regression tree is covered here. Roughly speaking, the idea in regression tree analysis is to divide the covariate space $\mathbb{X}^p$ in J distinct and non-overlapping regions, $R_1, \ldots, R_j$ (usually called *terminal nodes* or *leafs*) and to give the same predicted output value $\hat{c}_j$ for each observation in the same region R_j. Normally, this predicted value is the mean of the training samples output values in that region (or leaf) R_j. Regions $R_1, \ldots, R_j$ are multidimensional rectangles: in figure 2.2, a two dimensional covariate space is divided into five rectangles with the regression tree 2.1.

A tree is constructed with a *recursive binary splitting* approach. This is a *top-down*, *greedy* recursive partitioning algorithm. The term top-down means that the whole training data set is in one node called *root* at the beginning

(see example 2.1), and the algorithm goes forward by splitting the data in half at every node, until some predefined stopping criterion has been reached (for example, a minimum number of observations in a leaf). Greediness means that the best *binary split* (see Appendix 3) in that particular step is chosen without considering any further steps: the algorithm does no attempt to find the best binary split in a sense that it would optimize the whole tree with respect to the final residual sum of squares $\sum_{j=1}^{J}\sum_{i\in R_j}(y_i - \hat{y}_{R_j})^2$, where J is the total number of leafs, y_i is an observation in a training sample and $\hat{y}_{R_j}$ is the mean response for the training observations within the j^{th} box.

Minimizing the final residual sum of squares with the training data is not the desired end result: when the algorithm is used to prediction, one wants that the algorithm performs well beyond the training sample. If there are no constraints for the recursive binary splitting, the algorithm aims to fit the final tree T_0 perfectly in the training data. At the same time, the number of observations in terminal nodes decreases and thus the prediction variance increases, meaning that the model gets overfitted. It is possible to reduce model complexity, namely the size of the tree, during the tree construction process, for example by setting a minimum threshold value for the residual sum of squares that must be reduced in each step or the algorithm breaks. However, in this approach it is possible that a seemingly insignificant split may lead to better splits in the following steps. A more practical way is to construct a bigger tree T_0 without complexity restrictions and to find a less complex subtree by *pruning* the original tree. Pruning means that one finds a more optimal subtree $T^* \subset T_0$ with respect to bias-variance tradeoff. This can be done by validating the subtrees with a validation data set. However, it is computationally unfeasible to test all the possible subtrees. One strategy is to use *cost complexity pruning*: in this approach, one constructs a similar function for the tree model that was presented earlier for the LASSO-method (2.32):

$$\sum_{j=1}^{J}\sum_{i\in R_j}(y_i - \hat{y}_{R_j})^2 + \alpha|T|$$

In this function, term $\sum_{j=1}^{J}\sum_{i\in R_j}(y_i - \hat{y}_{R_j})^2$ is the residual sum of squares for subtree T and $\alpha|T|$ is the complexity penalty term with the number of terminal nodes $|T|$ and the tuning parameter α. An advantage of this function form is that when tuning parameter $\alpha > 0$, branches get pruned from the tree in a nested and predictable fashion [Breiman et al., 1984]. This makes it possible to obtain the whole sequence of subtrees as a function of α. Therefore, by selecting a tuning parameter α with the cross-validation (Appendix 4), one can select an optimal subtree T^*.

An advantage in decision trees is the interpretability of the results: especially with smaller trees, visualization of the results is easy and results are understandable even for a non-professional. Another advantage is that decision trees can handle all the basic data types: numeric, factors and binary data. However, decision trees are not particularly good predictors. The biggest reason for this

is that they suffer from a high variance [Hastie et al., 2009]. Secondly, the prediction surface has discrete jumps between the output regions and thus a terminal node mean can be biased for a data point lying far from a region center [Athey and Imbens, 2019]. These both problems can be managed by growing a large number of trees and then aggregating the results from individual trees. Aggregating the results from a large number of trees improves the prediction accuracy, yet at the same time it reduces the interpretability of the results.

One approach is *bagging*, in which multiple decision trees are grown by *bootstrapping* [Breiman, 1996]. Even if none of the individual trees is pruned, meaning that they have high variance but low bias, averaging over the individual predictions of the trees reduces the variance to a lower level as compared to a variance of a single decision tree. Nevertheless, the trees grown by bootstrapping are relatively similar to each other as far as the whole covariate space $\mathbb{X}$ is considered in each binary split. Therefore, the trees are highly correlated with each other and the variance cannot be reduced as much as with a lower correlation. In the random forest method, each split can only consider a randomized subsample of the whole covariate set [Breiman, 2001]. This reduces correlation between the trees and thus lowers the variance. Random forest has taken its place as one of the most popular machine learning algorithms [Athey and Imbens, 2019].

2.20 Causal Tree

In their article, Athey and Imbens [2016b] proposed a version of the regular decision tree (2.19) designed for the heterogeneous treatment effect estimation. The article considered two main topics: how two apply the decision tree algorithm to the problem of causal inference, and how to construct confidence intervals for the estimations. As mentioned above, the task of estimating treatment effects is harder than predicting outcomes because the treatment effect $\tau_i(x)$ is not observable for any individual, and thus the basic residual sum of squares cannot be directly applied in recursive splitting or in parameter tuning.

Secondly, the off-the-shelf regression tree methods cannot be used directly in estimation purposes because they are *adaptive*, meaning that the same data is used for the construction of trees and estimation. This leads to a bias in estimations that disappears slowly when the sample size increases. As an example, one can think of a splitting criterion, where the data is divided into two children $\{\{L, R\}\} \mapsto \{\{L\}, \{R\}\}$ if the subtraction of the means in the children leaves is greater than a threshold value c, $\bar{Y}_L - \bar{Y}_R \geq c$. Normally $\bar{Y}_L - \bar{Y}_R$ would be a unbiased estimator for $\mu(L) - \mu(R)$, but conditionally on $\bar{Y}_L - \bar{Y}_R \geq c$ this does not hold. For solving this problem, Athey and Imbens [2016b] proposed an *honest* estimation approach, in which half of the data from the training set S would be used in the tree construction (indicated as $\mathcal{J}$) and the other half in the terminal node estimation (indicated as $\mathcal{I}$). This is required in order to get unbiased estimates and for constructing confidence intervals. The cost of the honest estimation is the loss of statistical power when a smaller number of observations are used in estimation.

The aim in the causal tree algorithm is to find a partition in which the observations i falling into the same terminal nodes $i : X_i \in R$ would have similar covariate distributions with respect to important covariates according to treatment effect heterogeneity. Then one can estimate the conditional treatment effect $\hat{\tau}(x)$ in each terminal node by simply subtracting the conditional mean $\hat{\mu}(x, W_i = 0)$ from $\hat{\mu}(x, W_i = 1)$:

$$\begin{aligned}\hat{\tau}(x) = &\frac{1}{|\{i : W_i = 1, X_i \in R\}|} \sum_{\{i:W_i=1, X_i \in R\}} Y_i \\ &- \frac{1}{|\{i : W_i = 0, X_i \in R\}|} \sum_{\{i:W_i=0, X_i \in R\}} Y_i \end{aligned} \tag{2.33}$$

Here, the $|\{i : W_i = w, X_i \in R\}|$ represents the number of observations per group in a terminal node. [Wager and Athey, 2018]

The article by Athey and Imbens [2016b] provided several different spitting criteria for the causal tree method that are based on minimizing the estimated squared error loss in treatment effects. The follow-up paper Wager and Athey [2018] proposed a criterion that is based on the finding that maximizing the variance of $\hat{\tau}(x)$ is equivalent to minimizing the estimated squared-error :

$$\max_{(p,t)} \left\{ \sum_{i \in \mathcal{J}} \widehat{\mathbb{V}}(\hat{\tau}(X_i)) \right\}$$

The criterion is maximized with respect to covariate $p \in X$ and the cutpoint t. If a single causal tree were used in causal estimation, the tree would be pruned with cross-validation. Instead, if one uses multiple trees as discussed in subsections 2.19, there is no need for pruning.

2.21 Causal Forest

As for the basic decision tree, there exists a random forest approach for causal trees called causal forest, proposed originally by Wager and Athey [2018]. This method also uses the honest approach in treatment effect estimation, allowing the estimation of confidence intervals. Similarly as the random forest algorithm by Breiman [2001], the correlation between the individual causal trees is reduced by randomizing the subsamples of the data in tree construction and the possible set of covariates that the algorithm can consider in each split. In causal forest algorithm, the honest approach is more "efficient" with respect to the data as compared to its use in the construction of individual causal tree because each data point can be used in tree construction in some trees and in estimation in some other trees.

The causal forest method was synchronized into the class of generalized random forests in the paper by Athey et al. [2019]. This version of the method uses a gradient-based loss criterion in lieu of the exact loss criterion in the spirit of the gradient boosting algorithm by Friedman [2002], which makes the algorithm

computationally more efficient. Another difference between the methods is that the new version of the causal forest does not estimate the treatment effect by simply subtracting the conditional means in the leaves (2.33) and averaging the tree-specific results but by counting the forest-based weights. A weight $\alpha_i(x)$ represents the similarity of training example i with respect to the estimation target with the features x. The weight is constructed by counting the frequency of times when the i^{th} training observation falls into the same leaf as the target x. As it is discussed in subsection 2.18, the normal nearest-neighbor methods are not robust when the dimension of the covariate space increases. In contrast, when the kernel weights are derived in the forest-based way as in causal forest, the estimates are not as sensitive to the covariates that have only a little effect on treatment effects, making them more robust with a higher number of covariates.

To make the treatment effect estimates more robust on confounding, Athey et al. [2019] proposed to combine a method called *orthogonalization,* proposed originally by Robinson [1988], in the method. This means that before running the causal forest algorithm, one computes the propensity score estimates $\hat{e}(x) = \mathbb{E}(W_i = w \mid X_i)$ and marginal outcome estimates $\hat{m}(x) = \mathbb{E}(Y_i = y \mid X_i)$ by training separate random forests. Then the residual treatments $W_i - \hat{e}(X_i)$ and outcomes $Y_i - \hat{m}(X_i)$ are counted and used in the causal forest algorithm. In the simulation study 2.23, the causal forest methods with and without the orthogonalization are compared. The original paper by Robinson [1988] proposed an estimator for a partially linear model with a constant treatment effect, which was then later modified for heterogeneous treatment effect estimation [Nie and Wager, 2017].

By combining the methods mentioned above, one gets the CATE estimator for the causal forest [Athey and Wager, 2019]:

$$\hat{\tau}(x) = \frac{\sum_{i=1}^{n} \alpha_i(x) \left(Y_i - \hat{m}^{(-i)} X_i\right) \left(W_i - \hat{e}^{(-i)} X_i\right)}{\sum_{i=1}^{n} \alpha_i(x) \left(W_i - \hat{e}^{(-i)} X_i\right)^2} \quad \text{, where the subscript } {}^{(-i)}$$

denotes "out-of-bag" prediction. This means, that an outcome Y_i is not used to count $\hat{m}^{(-i)} X_i$ and an assignment observation W_i is not used to count $\hat{e}^{(-i)} X_i$. In their paper Athey et al. [2019] showed that the estimates of the causal forest are consistent and asymptotically Gaussian, and they provided an estimator for the asymptotic variance that enables valid confidence intervals. However, many empirical studies, including a simulation study in the original paper [Athey et al., 2019] and the simulation study of this thesis (2.23), have showed that in many scenarios the coverage of the confidence intervals is still unsatisfying.

2.22 Earlier Treatment Effect Estimation Simulation Studies

Due to the "fundamental problem of causal inference" (2.2), methodological papers considering treatment effect estimation have traditionally favored simulation studies. This is especially true for the papers considering heterogeneous

treatment effect estimation [e.g. Hill, 2011, Athey and Imbens, 2016b, Hahn et al., 2017, Powers et al., 2017, Atan et al., 2018, Carvalho et al., 2019] and observational data estimation methods [e.g. Maldonado and Greenland, 1993, MacKinnon et al., 1995, Fewell et al., 2007], where RCT results would either not reveal the true underlying heterogeneity or not have the assignment mechanisms of interest. In this thesis in which the topic is heterogeneous treatment effect estimation from the observational data, neither of these subjects would be empirically covered. Nevertheless, some papers have tried to mix observational studies with RCT data. An article by Shadish and Cook [2009] compared results estimated with propensity score methods from observational data to the corresponding RCT results with an aim to asses credibility for the methods. However, this approach was criticized by Pearl [2009c]: propensity score methods are sensitive to covariate selection, and the fact that the method works for a particular dataset does not give external validity for the method.

One way to execute an empirical study is to generate semi-simulated data. Hill [2011], Johansson et al. [2016], Atan et al. [2018] used data from an RCT that evaluated the impact of the *Impact of the Infant Health and Development Program* on the subjects' IQ scores at the age of three [Brooks-Gunn et al., 1992, Bradley et al., 1994], from which a subset of the treated population (all children with non-white mothers) was removed to introduce selection bias. The two outcome surfaces were generated with the RCT covariates so that unconfoundedness and overlap assumptions were filled and the true treatment effects were known. Wendling et al. [2018] compared heterogeneous treatment effect estimation methods with health care database data where real covariate and treatment assignment data were used and only outcomes were simulated based on nonparametric models of the real outcomes.

Similarly, in the work shop *2018 Atlantic Causal Inference Conference* eight groups of researchers analyzed a synthetic observational dataset that was generated using the statistical results from a recent large-scale RCT in education Carvalho et al. [2019]. The idea was to create an observational dataset with the original covariate distributions, data structures, and effect sizes, but where it was possible, researchers added a synthetic data generating-process to the original data [Yeager et al., 2019] and fitted semiparametric models to the post-treatment outcomes from the original study. The confounding was generated in the data by a two-step process: at first, observations were dropped with probability $1 - \Phi(-0.5 + 1.5\mu(w_{ij}))$, which simulates *"a scenario where students with high expected outcomes under control were more likely to receive the treatment, yielding naive treatment effect estimates that are too high"*. Then, selected units from the treatment arm were dropped to induce a more complicated functional form for the confounding structure.

In the simulation study of this thesis, data is created by defining underlying SCMs and then generating it with a *Monte Carlo simulation*. The same approach was used in an article written by Liu et al. [2012] where small sample properties of the M-structure (2.16) were studied. In the study, data for independent variables was generated independently by using results from the existing literature related to substance [McCall et al., 2002, Arday et al., 2002],

and the data for remaining variables was structurally derived from the independent variables, by using parameters that are based on epidemiological literature [Levy, 1981, Wilson et al., 1988, Friedman et al., 1997, Wilson et al., 1998, Federman et al., 2001, Arday et al., 2002, McCall et al., 2002, Crystal et al., 2003, Ford et al., 2003, Wulsin and Singal, 2003, Ma et al., 2005]. In the simulation, binary variables were simulated with logistic regression models and continuous variables with log-linear models.

2.23 Simulation Study

The simulation study is done with the `R`-program (information about the used codes and packages in Appendix K). The data is generated based on the SCMs that are presented in the following nine subsections 2.23.1–2.23.9. The treatment effects are estimated by fitting different kinds of causal forest models for each nine datasets. The aim is to compare estimation accuracy with models that differ in the covariate sets offered for the adjusting in different kinds of causal structures. Also, the orthogonalization is tested with SCMs 2.23.3 and 2.23.7. The prediction accuracy is measured with the *root mean squared error* (RMSE)

$$RMSE\left(\hat{\tau}(x)\right) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\tau}_i(x) - \tau_i(x)\right)^2}$$

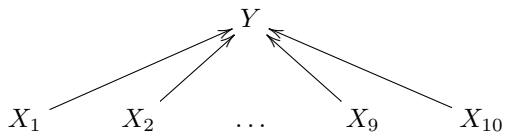
and the *95 percent confidence interval coverage*. The inner function of the RMSE, MSE $\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\tau}_i(x) - \tau_i(x)\right)^2$, is also called the *reducible error*, which is the other component of the *expected prediction error* alongside the *irreducible error* $\mathbb{V}_{\tau|X}[\tau \mid X = x]$. As the name suggests, the reducible error part is the one over which one has some control with model selection, whereas the irreducible error is the noise in the parameter of interest, namely the part that one does not like to learn. The mean squared error can be decomposed forward into to parts: square of the bias of an estimator and variance of the estimator:

$$\mathrm{MSE}\left(\tau(x), \hat{\tau}(x)\right) = \underbrace{\left(\tau(x) - \mathbb{E}\left[\hat{\tau}(x)\right]\right)^2}_{\mathrm{bias}^2(\hat{\tau}(x))} + \underbrace{\mathbb{E}\left[\left(\hat{\tau}(x) - \mathbb{E}\left[\hat{\tau}(x)\right]\right)^2\right]}_{\mathbb{V}(\hat{\tau}(x))}$$

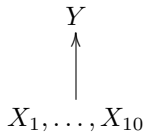
This is the formalization of the bias-variance tradeoff that is mentioned in subsection 2.18. The reason why RMSE is used instead of MSE is that the root operator makes the measure based on the same units as the quantity being estimated, increasing the interpretability of the results. The 95 percent confidence interval coverage is provided to test the robustness of the standard deviation estimates.

The DAGs in the following subsections use the following notations: the group of observed variables is $\{W, Y, \mathbb{X}, \mathbb{Z}, \mathbb{C}, M\}$ and unobserved $\{\mathbb{U}\}$. Variables

$\{\mathbb{X}, \mathbb{Z}, \mathbb{C}, \mathbb{U}\}$ are assumed to be exogenous. For simplicity, the DAG structure



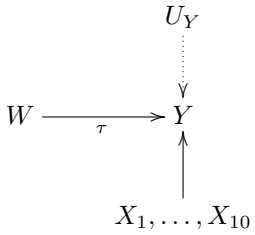
is indicated as the following:



Simulations 2.23.1 and 2.23.6 have random assignments, simulation 2.23.2 has a non-random but unconfounded assignment, and the remaining ones have confounded assignments. Simulations 2.23.1–2.23.5 have a constant treatment effect $\tau = 10$, and 2.23.6–2.23.9 heterogeneous treatment effects. The parameter values of the simulation study can be found in Appendix J. In each simulation, the sample size is $N = 10,000$. The output functions f_Y are generated with linear models in each simulation, and the treatment assignment functions f_W with logistic functions in each non-random case. The overlapping assumption 2.7 is filled in every simulation.

2.23.1 Randomized Controlled Trial with Constant Treatment Effect, Including Covariates Affecting the Output

The first simulation is an RCT with an assignment variable W, an exogenous unobserved variable U_Y, and ten observed covariates $X_1, \ldots, X_{10}$ that affect the outcome Y:



$$f_W = \text{Bern}\left(\frac{1}{2}\right)$$

$$f_Y(W, X_1, \ldots, X_{10}, U_Y) = \tau \cdot W + \sum_{k=1}^{10} \beta_{X_k} X_k + U_Y \qquad \text{, where}$$

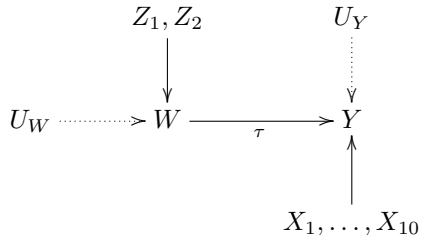
$$\begin{cases} U_Y & \sim \text{Norm}(\mu_Y(0), \sigma_Y^2) \\ X_K & \sim \text{Norm}(\mu_K, \sigma_K^2) \text{ when } k \text{ is odd and} \\ X_K & \sim \text{Bern}\left(\frac{1}{2}\right) \text{ when } k \text{ is even} \end{cases}$$

Two causal forest models are compared this simulation: one with the whole set of observed covariates in the adjusted set $\{X_1, \ldots, X_{10}\} = \mathbb{Z}_{\text{full}}$, and another with a subset $\{X_1, X_3, X_5, X_9\} = \mathbb{Z}_{\text{subset}}$. The latter adjusted set $\mathbb{Z}_{\text{subset}}$ is chosen by selecting the four most important variables from the set $\mathbb{Z}_{\text{full}}$ when the importance is measured by a simple weighted sum of how many times feature X_p was split at each depth in the forest.

The distributions for the predicted treatment effects for the models are plotted on figure 2.3. The estimated ATEs are approximately the same for both of the models, 10.23 with standard deviations 0.20. As regards to this, it seems likely that both of the models are upwardly biased (p-value for ATE $\mathbb{E}(\tau) = 10$ is under five percent). As it can be seen from table 1, the full model performs slightly better than the subset model. Both of the coverages are over 95 percent.

2.23.2 Constant Treatment Effect with Unconfounded Assignment, Including Covariates Affecting the Output

The second simulation is similar to the first one with the difference that the treatment assignment is not random. Still, the causal model is unconfounded:



$$f_W(Z_1, Z_2, U_W) = \frac{1}{1 + \exp\left\{-\left(\sum_{j=1}^{2} \gamma_{Z_j} Z_j + U_W\right)\right\}}$$

$$f_Y(W, X_1, \ldots, X_{10}, U_Y) = \tau \cdot W + \sum_{k=1}^{10} \beta_{X_k} X_k + U_Y \qquad \text{, where}$$

$$\begin{cases} U_W & \sim \text{Norm}(0, \sigma_W^2) \\ U_Y & \sim \text{Norm}(\mu_Y(0), \sigma_Y^2) \\ Z_1, Z_2 & \sim \text{Bern}\left(\frac{1}{2}\right) \\ X_K & \sim \text{Norm}(\mu_K, \sigma_K^2) \text{ when } k \text{ is odd and} \\ X_K & \sim \text{Bern}\left(\frac{1}{2}\right) \text{ when } k \text{ is even} \end{cases}$$

For this model, two forest models were fitted: one with all the observed variables $\{Z_1, Z_2, X_1, \ldots, X_{10}\} = \mathbb{Z}_{\text{full}}$ and the other excluding the observed

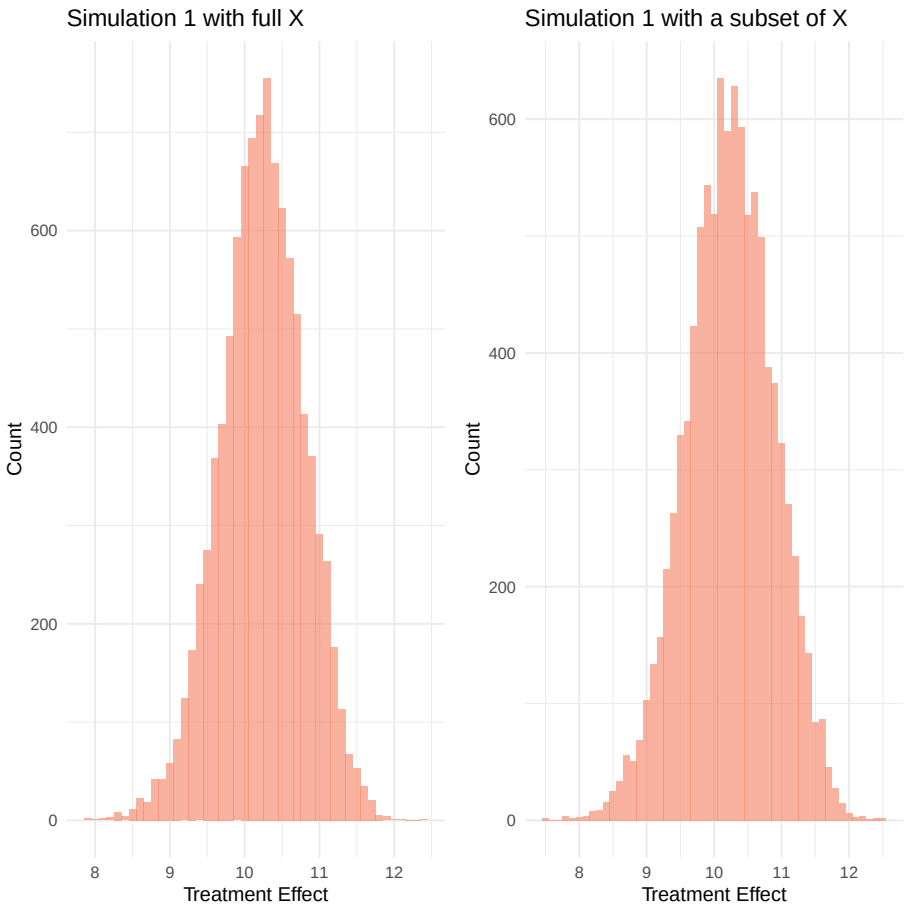


Figure 2.3: Simulation 1 predicted treatment effects.

	RMSE
Full X	0.6108580
Subset of X	0.6979229

	Coverage
Full X	0.9917
Subset of X	0.9875

Table 1: RMSEs and coverages in simulation 1.

	RMSE
Without Z	0.6930577
With Z	0.6742028

	Coverage
Without Z	0.9842
With Z	0.9877

Table 2: RMSEs and coverages in simulation 2.

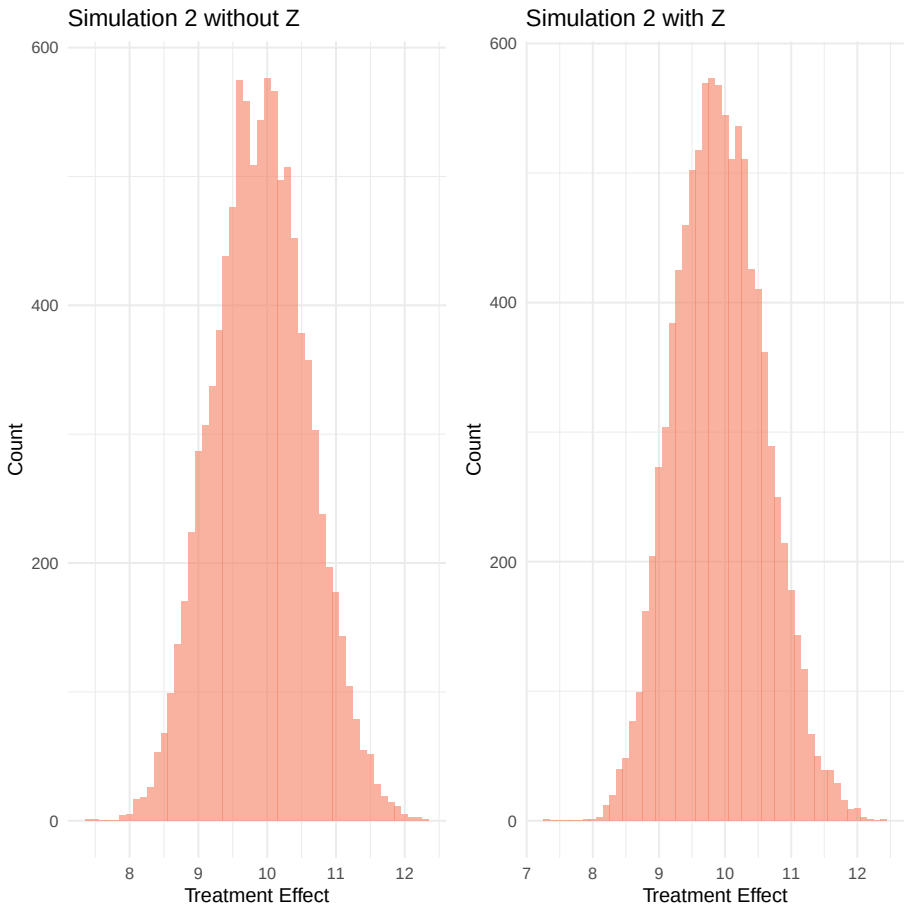
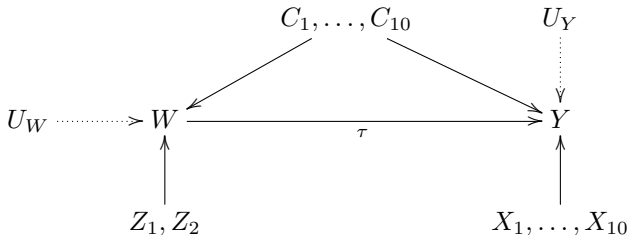


Figure 2.4: Simulation 2 predicted treatment effects.

parents of the assignment variable $\{X_1, \ldots, X_{10}\} = \mathbb{Z}_{\text{subset}}$. Asymptotically both of the adjusted sets fill the unconfoundedness criterion. In this case, the causal forest that considers all the observed variables gives a slightly lower RMSE, as can be seen from table 2. In this case, both of the models estimate the ATE relatively accurately with approximately the same estimations 9.93 with the standard deviations 0.20.

2.23.3 Constant Treatment Effect with Confounded Assignment, Including Covariates Affecting the Output

In addition to the variables of the previous simulation, this simulation has ten confounders $\{C_1, \ldots, C_{10}\}$:



$$f_W(Z_1, Z_2, C_1, \ldots, C_{10}, U_W) = \frac{1}{1 + \exp\left\{-\left(\sum_{j=1}^{2} \gamma_{Z_j} Z_j + \sum_{l=1}^{10} \gamma_{C_l} C_l + U_Y\right)\right\}}$$

$$f_Y(W, X_1, \ldots, X_{10}, C_1, \ldots, C_{10}, U_Y) = \tau \cdot W + \sum_{k=1}^{10} \beta_{X_k} X_k + \sum_{l=1}^{10} \beta_{C_l} C_l + U_Y \quad , \text{ where}$$

$$\begin{cases} U_Y & \sim \text{Norm}(\mu_Y(0), \sigma_Y^2) \\ U_W & \sim \text{Norm}(0, \sigma_W^2) \\ Z_1, Z_2 & \sim \text{Bern}\left(\frac{1}{2}\right) \\ X_K & \sim \text{Norm}(\mu_K, \sigma_K^2) \text{ when } k \text{ is odd and} \\ X_K & \sim \text{Bern}\left(\frac{1}{2}\right) \text{ when } k \text{ is even} \\ C_L & \sim \text{Norm}(\mu_L, \sigma_L^2) \text{ when } l \text{ is odd and} \\ C_L & \sim \text{Bern}\left(\frac{1}{2}\right) \text{ when } l \text{ is even} \end{cases}$$

In this case, four different models are tested: the first with the whole set of observed variables $\{C_1, \ldots, C_{10}, Z_1, Z_2, X_1, \ldots, X_{10}\} = \mathbb{Z}_{\text{full}}$ and without orthogonalization, the second with the whole set of observed variables $\mathbb{Z}_{\text{full}}$ and with orthogonalization, the third with a subset of "important variables" $\mathbb{Z}_{\text{subset}}$ chosen similarly to the first simulation (2.23.1) and with orthogonalization, and the last with the set of confounders $\{C_1, \ldots, C_{10}\} = \mathbb{Z}_{\text{back-door}}$ adjusted and with orthogonalization. As regards to the back-door criterion (2.14), set $\mathbb{Z}_{\text{back-door}}$ is the minimum set of variables for filling the unconfoundedness assumption.

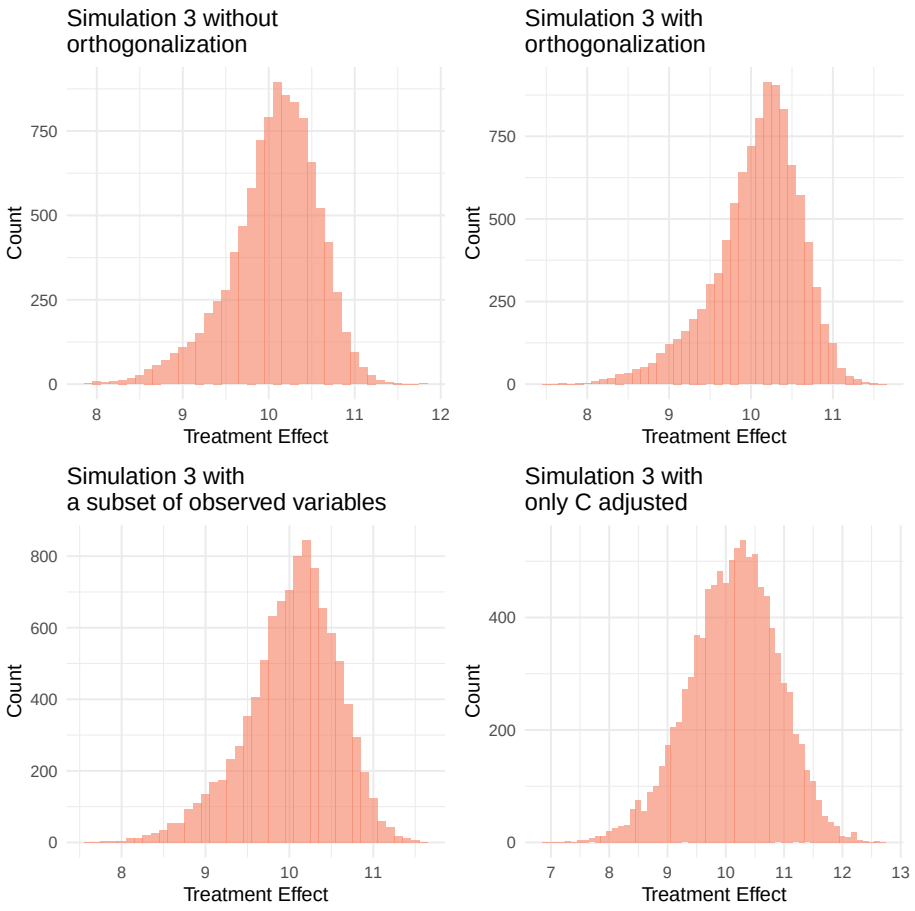


Figure 2.5: Simulation 3 predicted treatment effects.

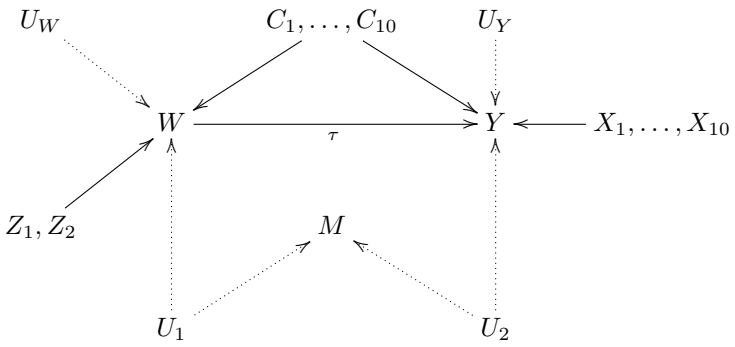
	RMSE
No orthogonalization	0.5160037
With orthogonalization	0.5342760
Subset of observed variables	0.5526428
Only confounders adjusted	0.7784988
	Coverage
No orthogonalization	0.9976
With orthogonalization	0.9966
Subset of observed variables	0.9972
Only confounders adjusted	0.9815

Table 3: RMSEs and coverages in simulation 3.

With respect to RMSE, the performances of the models are in the same order as they are listed above (table 3). The first three of the models do not have significant differences in their estimation accuracy, but the model with adjusted set $\mathbb{Z}_{\text{back-door}}$ performs significantly worse. Also the estimated ATE for the last model differs from the true treatment effect 1.3 percent when the difference for the other models is below 0.5 percent. It is also an interesting result that the orthogonalization does not improve the performance in the case of the confounded assignment.

2.23.4 Constant Treatment Effect with Confounded Assignment, Including Covariates Affecting the Output and a Pure Local M-Structure

The fourth simulation is built over the simulation three with a local M-structure $W \dashleftarrow U_1 \dashrightarrow M \dashleftarrow U_2 \dashrightarrow Y$:



$$f_W(Z_1, Z_2, C_1, \ldots, C_{10}, U_1, U_W) = \frac{1}{1 + \exp\left\{-\left(\sum_{j=1}^{2} \gamma_{Z_j} Z_j + \sum_{l=1}^{10} \gamma_{C_l} C_l + \gamma_{U_1} U_1 + U_Y\right)\right\}}$$

$$f_M = \delta_1 U_1 + \delta_2 U_2$$

$$f_Y(W, X_1, \ldots, X_{10}, C_1, \ldots, C_{10}, I, U_2, U_Y) = \tau \cdot W + \sum_{k=1}^{10} \beta_{X_k} X_k + \sum_{l=1}^{10} \beta_{C_l} C_l + \beta_{U_2} U_2 + U_Y \quad , \text{ where}$$

$$\begin{cases} U_Y & \sim \text{Norm}(\mu_Y(0), \sigma_Y^2) \\ U_W & \sim \text{Norm}(0, \sigma_W^2) \\ Z_1, Z_2 & \sim \text{Bern}\left(\frac{1}{2}\right) \\ X_K & \sim \text{Norm}(\mu_K, \sigma_K^2) \text{ when } k \text{ is odd and} \\ X_K & \sim \text{Bern}\left(\frac{1}{2}\right) \text{ when } k \text{ is even} \\ C_L & \sim \text{Norm}(\mu_L, \sigma_L^2) \text{ when } l \text{ is odd and} \\ C_L & \sim \text{Bern}\left(\frac{1}{2}\right) \text{ when } l \text{ is even} \\ U_1, U_2 & \sim \text{Bern}\left(\frac{1}{2}\right) \end{cases}$$

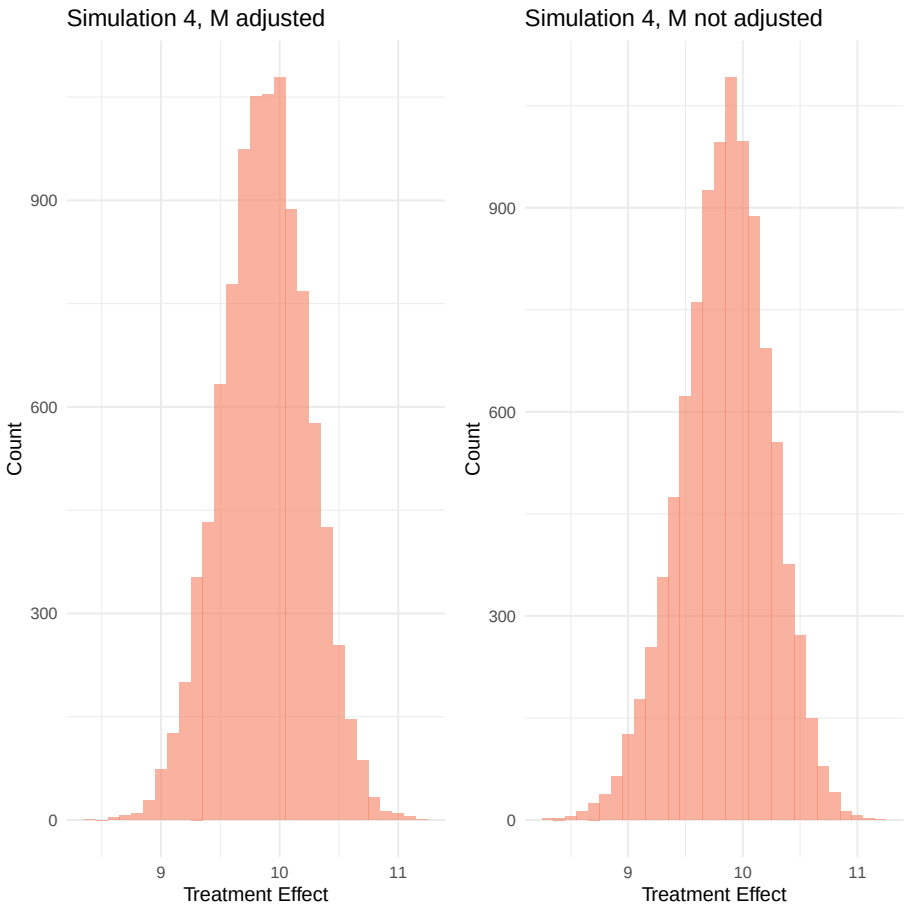


Figure 2.6: Simulation 4 predicted treatment effects.

	RMSE
M adjusted	0.3844006
M not adjusted	0.4186501

	Coverage
M adjusted	0.9998
M not adjusted	0.9994

Table 4: RMSEs and coverages in simulation 4.

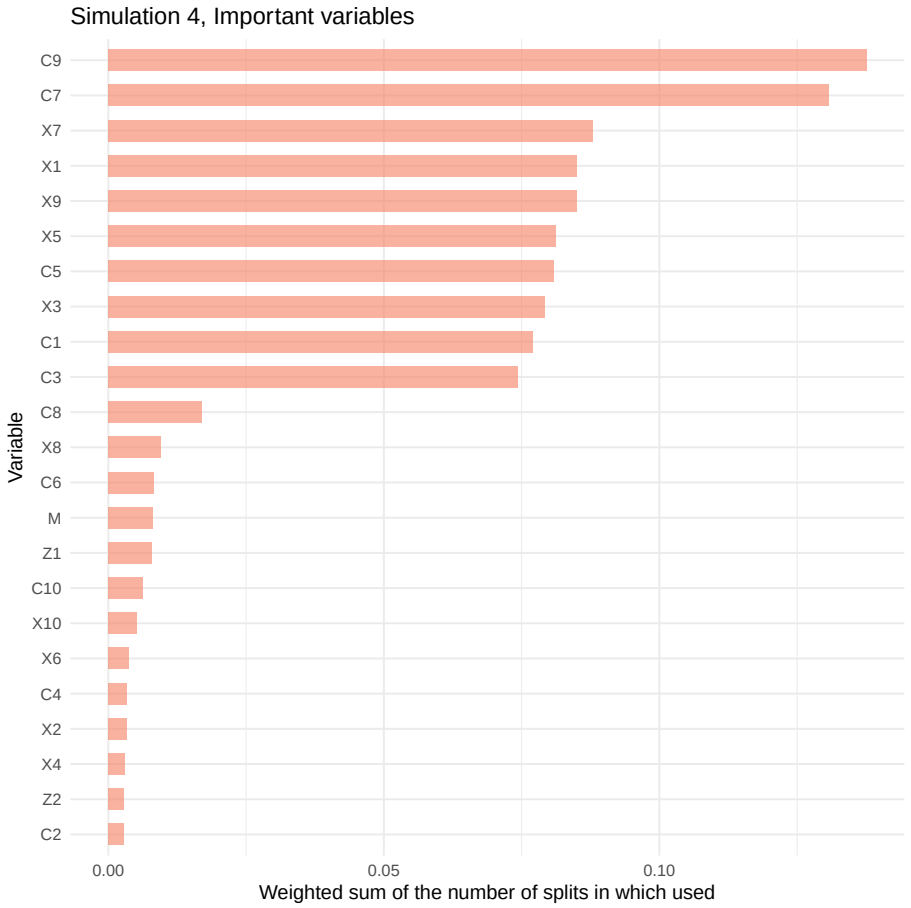
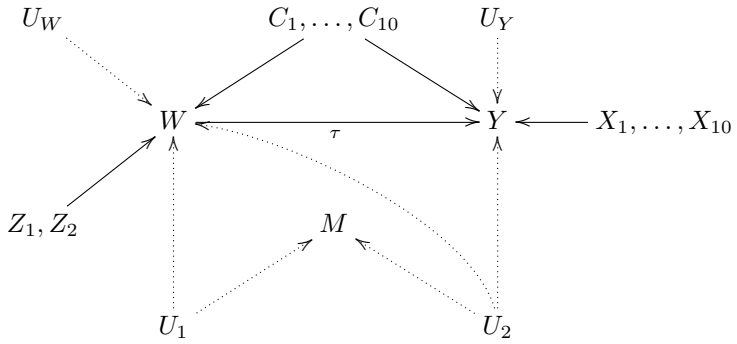


Figure 2.7: Simulation 4 important variables, where importance means a weighted sum of how many times a feature was split on at each depth in the forest.

Discussion about the M-structure is provided in subsection 2.16. Here, the model comparison is done between two causal forest models. The other model considers the full set of observed variables, including the variable M ($\mathbb{Z}_{\text{IncM}}$), whereas in the other model variable M is not conditioned ($\mathbb{Z}_{\text{ExcM}}$). As mentioned in subsection 2.16, adjusting the collider in a M-structure should import bias in the estimates. Interestingly, this was not the case in this simulation: as can be seen from table 4, the model with the adjusted M provides higher accuracy with respect to the RMSE with approximately full coverage. Both of the models provide a relatively good estimation of ATE: model $\mathbb{Z}_{\text{IncM}}$ 9.87 (0.22) and $\mathbb{Z}_{\text{ExcM}}$ 9.85 (0.22). The importance of variable M as compared to the other variables can be seen in figure 2.7. As one can see, the importance of the variable M is relatively low.

2.23.5 Constant Treatment Effect with Confounded Assignment, Including Covariates Affecting the Output and an Impure Local M-Structure

The last simulation with the constant treatment effect $\tau = 10$ is similar to the previous one but with a directed edge $U_2 \dashrightarrow W$ breaking the local pure M-structure. This is one scenario that Imbens [2019] mentioned in his essay:



$$f_W(Z_1, Z_2, C_1, \ldots, C_{10}, U_1, U_2, U_W) = \frac{1}{1 + \exp\left\{-\left(\sum_{j=1}^{2} \gamma_{Z_j} Z_j + \sum_{l=1}^{10} \gamma_{C_l} C_l + \sum_{m=1}^{2} \gamma_{U_m} U_m + U_Y\right)\right\}}$$

$$f_M = \delta_1 U_1 + \delta_2 U_2$$

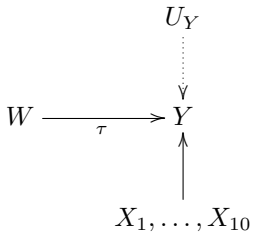
$$f_Y(W, X_1, \ldots, X_{10}, C_1, \ldots, C_{10}, U_2, U_Y) = \tau \cdot W + \sum_{k=1}^{10} \beta_{X_k} X_k + \sum_{l=1}^{10} \beta_{C_l} C_l + \beta_{U_2} U_2 + U_Y \quad , \text{ where}$$

$$\begin{cases} U_Y & \sim \text{Norm}(\mu_Y(0), \sigma_Y^2) \\ U_W & \sim \text{Norm}(0, \sigma_W^2) \\ Z_1, Z_2 & \sim \text{Bern}\left(\frac{1}{2}\right) \\ X_K & \sim \text{Norm}(\mu_K, \sigma_K^2) \text{ when } k \text{ is odd and} \\ X_K & \sim \text{Bern}\left(\frac{1}{2}\right) \text{ when } k \text{ is even} \\ C_L & \sim \text{Norm}(\mu_L, \sigma_L^2) \text{ when } l \text{ is odd and} \\ C_L & \sim \text{Bern}\left(\frac{1}{2}\right) \text{ when } l \text{ is even} \\ U_1, U_2 & \sim \text{Bern}\left(\frac{1}{2}\right) \end{cases}$$

Subsection 2.16 discusses this structure. The results of this simulation give similar results as previous literature that suggests conditioning the collider M in this impure M-structure case [Ding and Miratrix, 2015]. As can be seen from table 5, the RMSE of the model with the adjusted set $\mathbb{Z}_{\text{IncM}}$ is slightly below the other model with $\mathbb{Z}_{\text{ExcM}}$. Nevertheless, the RMSEs and coverages are nearly identical in the two models, as are the ATE estimates (10.05 (0.22) for both of the models).

2.23.6 Randomized Controlled Trial with Heterogeneous Treatment Effect, Including Covariates Affecting the Output

All the remaining simulations have heterogeneous treatment effects. Apart from the treatment effect, this simulation corresponds to simulation 2.23.1:



$$f_W = \text{Bern}\left(\frac{1}{2}\right)$$

$$f_Y(W, X_1, \ldots, X_{10}, U_Y) = \tau(x_1, x_2, x_3) \cdot W + \sum_{k=1}^{10} \beta_{X_k} X_k + U_Y \qquad \text{, where}$$

$$\begin{cases} U_Y & \sim \text{Norm}(\mu_Y(0), \sigma_Y^2) \\ X_K & \sim \text{Norm}(\mu_K, \sigma_K^2) \text{ when } k \text{ is odd and} \\ X_K & \sim \text{Bern}\left(\frac{1}{2}\right) \text{ when } k \text{ is even} \end{cases}$$

and $\tau(x_1, x_2, x_3) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 x_3$.

As done in simulation 2.23.1, two causal forest models are estimated: one with a full set of variables $\mathbb{Z}_{\text{full}}$, and one with a subset of important variables, $\mathbb{Z}_{\text{subset}}$, selected from the $\mathbb{Z}_{\text{full}}$ model as earlier. Not surprisingly, in the case

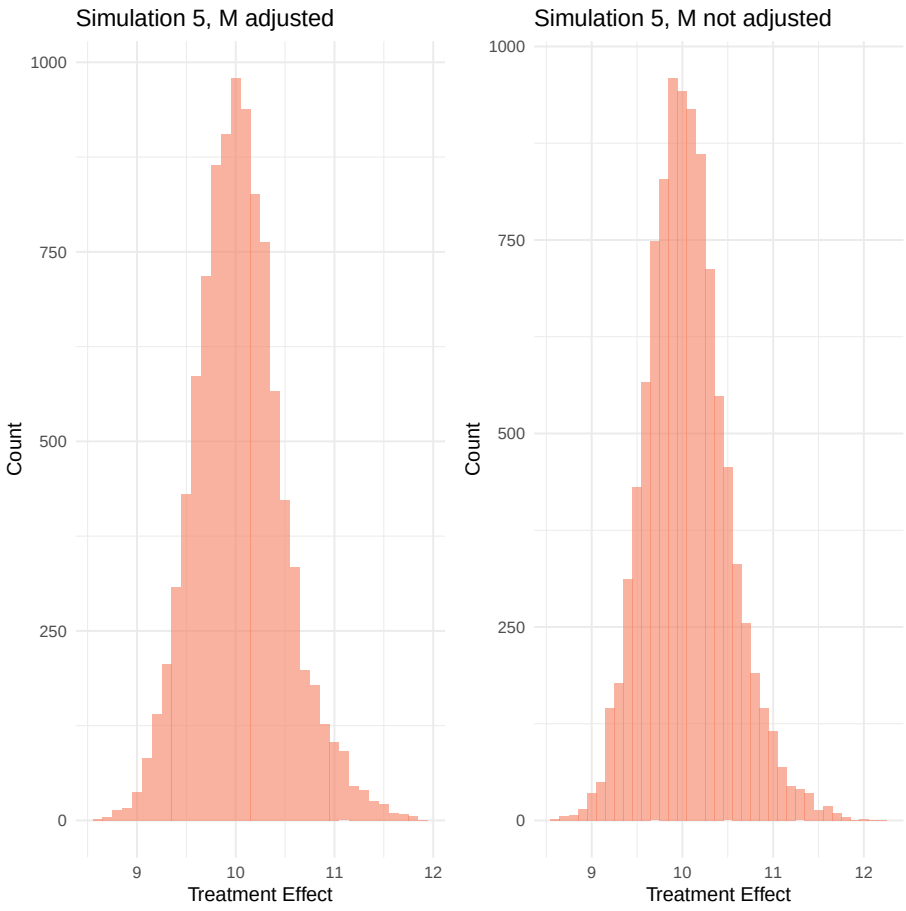


Figure 2.8: Simulation 5 predicted treatment effects.

	RMSE
M adjusted	0.4490187
M not adjusted	0.4513016

	Coverage
M adjusted	0.9996
M not adjusted	0.9986

Table 5: RMSEs and coverages in simulation 5.

	RMSE
Full set X	2.274664
Subset of X	2.877371

	Coverage
Full set X	0.7734
Subset of X	0.6238

Table 6: RMSEs and coverages in simulation 6.

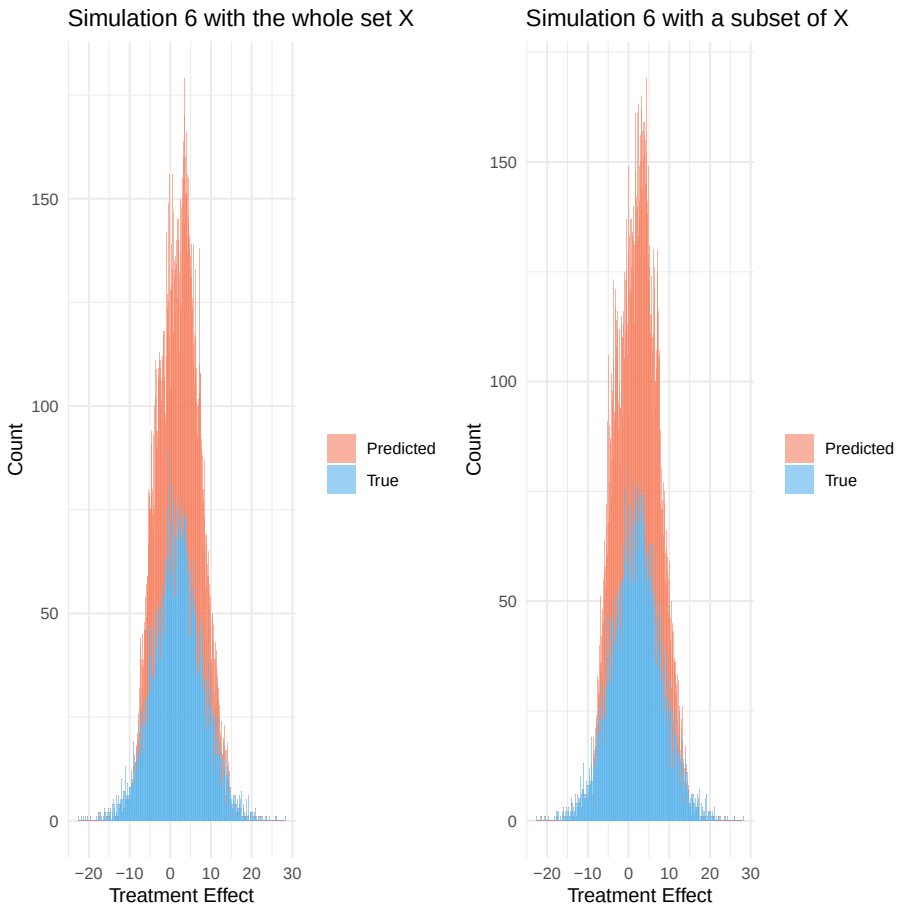


Figure 2.9: Simulation 6 predicted and true treatment effects.

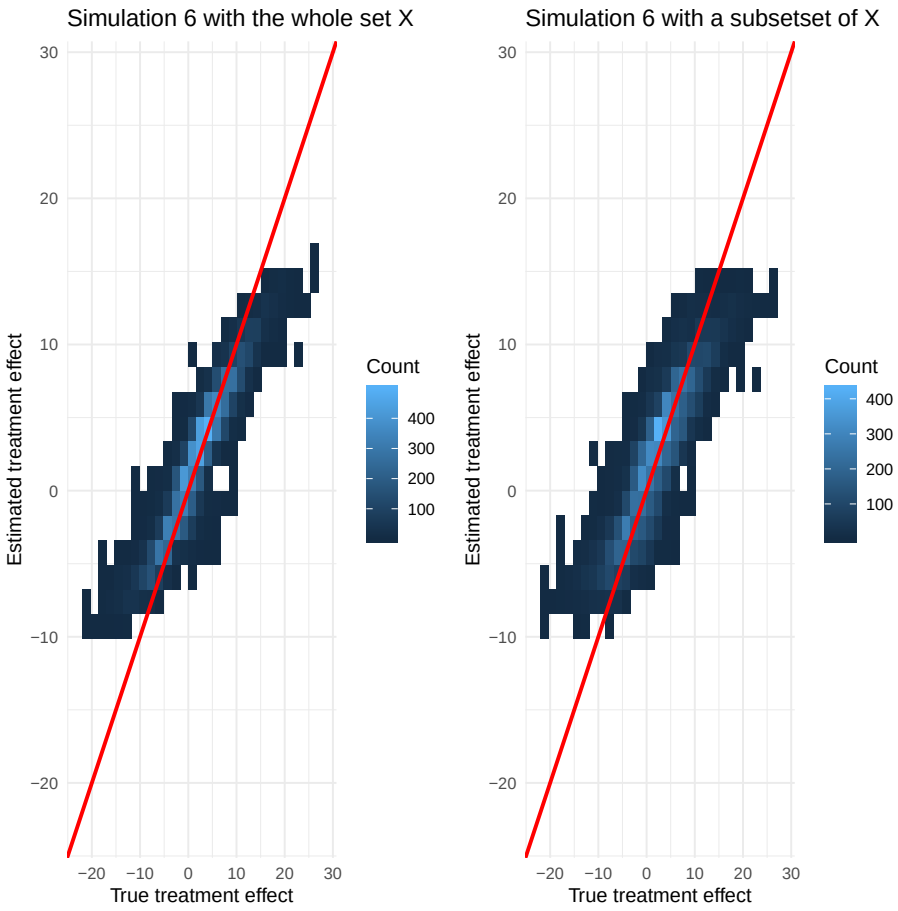


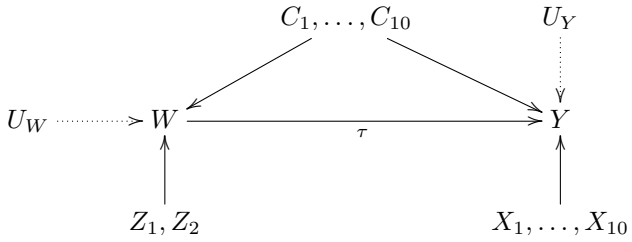
Figure 2.10: Simulation 6 predicted and true treatment effects.

of heterogeneous treatment effect estimation, the model with the full set of adjusted variables made significantly better estimations than subset model 6. One probable reason for this is the fact that the chosen subset does not include variable X_2 which is part of the treatment effect function: the way in which the important variables are selected favors the continuous variables over the binary (more potential splits), even if some binary variables have a great impact on the true underlying effect.

From figures 2.9 and 2.10 one can notice that both of the models provide too conservative estimates for the observations that have a true treatment further from the population ATE. One important point is that the coverage rates have fallen greatly from those that were estimated in the constant treatment effect cases. More importantly, estimated 95 percent confidence interval coverage rates are notably under 95 percent, reducing the creditably of the robustness of the estimated standard deviations.

2.23.7 Heterogeneous Treatment Effect with Confounded Assignment, Including Covariates Affecting the Output

This simulation is based on the same SCM as 2.23.3 with the heterogeneous treatment effect function $\tau(x_1, x_2, x_3, c_1)$:



$$f_W(Z_1, Z_2, C_1, \ldots, C_{10}, U_W) = \frac{1}{1 + \exp\left\{-\left(\sum_{j=1}^{2} \gamma_{Z_j} Z_j + \sum_{l=1}^{10} \gamma_{C_l} C_l + U_Y\right)\right\}}$$

$$f_Y(W, X_1, \ldots, X_{10}, C_1, \ldots, C_{10}, U_Y) = \tau(x_1, x_2, x_3, c_1) \cdot W + \sum_{k=1}^{10} \beta_{X_k} X_k + \sum_{l=1}^{10} \beta_{C_l} C_l + U_Y \quad , \text{ where}$$

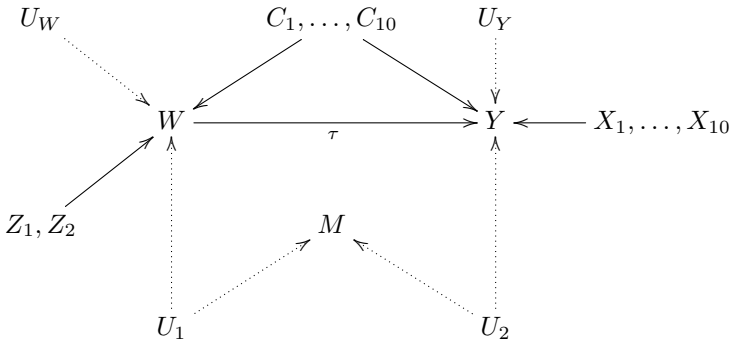
$$\begin{cases} U_Y & \sim \text{Norm}(\mu_Y(0), \sigma_Y^2) \\ U_W & \sim \text{Norm}(0, \sigma_W^2) \\ Z_1, Z_2 & \sim \text{Bern}\left(\frac{1}{2}\right) \\ X_K & \sim \text{Norm}(\mu_K, \sigma_K^2) \text{ when } k \text{ is odd and} \\ X_K & \sim \text{Bern}\left(\frac{1}{2}\right) \text{ when } k \text{ is even} \\ C_L & \sim \text{Norm}(\mu_L, \sigma_L^2) \text{ when } l \text{ is odd and} \\ C_L & \sim \text{Bern}\left(\frac{1}{2}\right) \text{ when } l \text{ is even} \end{cases}$$

and $\tau(x_1, x_2, x_3, c_1) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 x_3 + \alpha_3 c_1$.

Two causal forest models were tested for this data: one with orthogonalization and one without. As earlier in simulation 2.23.3, the orthogonalization does not improve the performance. However, it does not significantly impair it either. As in the previous simulation 2.23.6, the estimated treatment effects are too conservative for the tails of the true treatment effects. For both of the models, coverage stays below 70 percent, as one can see from table 7.

2.23.8 Heterogeneous Treatment Effect with Confounded Assignment, Including Covariates Affecting the Output and a Pure Local M-Structure

The eighth simulation is the heterogeneous treatment effect version of simulation 2.23.4:



$$f_W(Z_1, Z_2, C_1, \ldots, C_{10}, U_1, U_W) = \frac{1}{1 + \exp\left\{-\left(\sum_{j=1}^{2} \gamma_{Z_j} Z_j + \sum_{l=1}^{10} \gamma_{C_l} C_l + \gamma_{U_1} U_1 + U_Y\right)\right\}}$$

$$f_M = \delta_1 U_1 + \delta_2 U_2$$

$$f_Y(W, X_1, \ldots, X_{10}, C_1, \ldots, C_{10}, U_2, U_Y) = \tau(x_1, x_2, x_3, c_1) \cdot W + \sum_{k=1}^{10} \beta_{X_k} X_k + \sum_{l=1}^{10} \beta_{C_l} C_l + \beta_{U_2} U_2 + U_Y \quad , \text{ where}$$

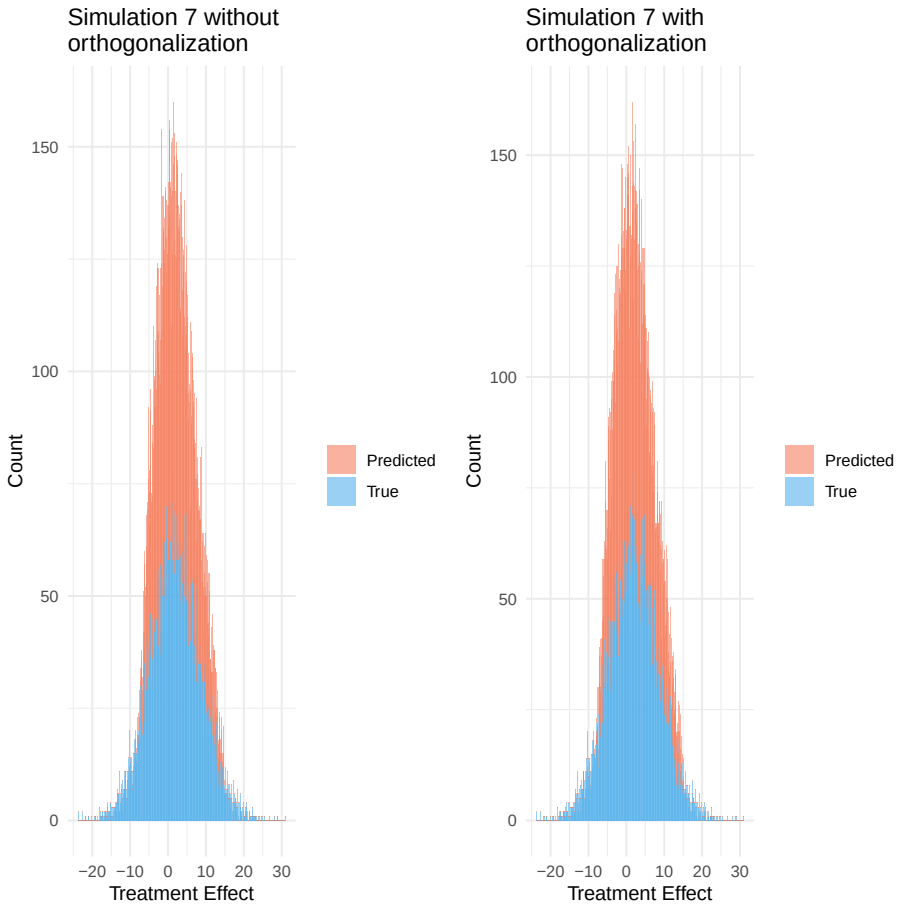


Figure 2.11: Simulation 7 predicted and true treatment effects.

	RMSE
With orthogonalization	3.157229
Without orthogonalization	3.101125
	Coverage
With orthogonalization	0.6832
Without orthogonalization	0.6911

Table 7: RMSEs and coverages in simulation 7.

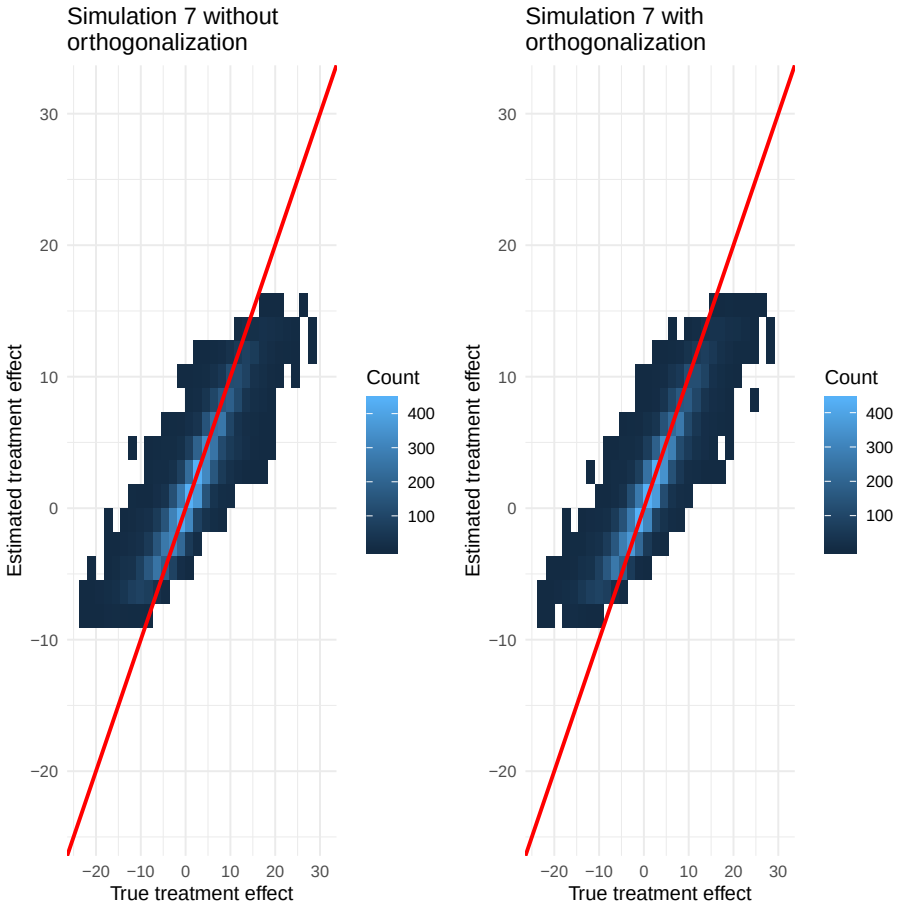


Figure 2.12: Simulation 7 predicted and true treatment effects.

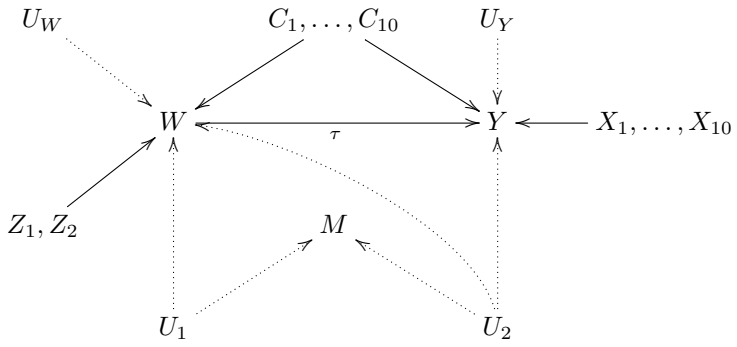
$$\begin{cases} U_Y & \sim \text{Norm}(\mu_Y(0), \sigma_Y^2) \\ U_W & \sim \text{Norm}(0, \sigma_W^2) \\ Z_1, Z_2 & \sim \text{Bern}\left(\frac{1}{2}\right) \\ X_K & \sim \text{Norm}(\mu_K, \sigma_K^2) \text{ when } k \text{ is odd and} \\ X_K & \sim \text{Bern}\left(\frac{1}{2}\right) \text{ when } k \text{ is even} \\ C_L & \sim \text{Norm}(\mu_L, \sigma_L^2) \text{ when } l \text{ is odd and} \\ C_L & \sim \text{Bern}\left(\frac{1}{2}\right) \text{ when } l \text{ is even} \\ U_1, U_2 & \sim \text{Bern}\left(\frac{1}{2}\right) \end{cases}$$

and $\tau(x_1, x_2, x_3, c_1) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 x_3 + \alpha_3 c_1$.

As in simulation 2.23.4, one of the two estimated models has a collider M in the adjusted set $\mathbb{Z}_{\text{IncM}}$ and the other one does not. From a theoretical point of view, it is interesting to see that excluding the M does not improve the estimation but seems to even weaken it marginally. Still the coverage rates are dissatisfying (table 8) and treatment effect estimates for the tails are too conservative (figures 2.13 and 2.14).

2.23.9 Heterogeneous Treatment Effect with Confounded Assignment, Including Covariates Affecting the Output and an Impure Local M-Structure

The last simulation is the heterogeneous treatment effect version of simulation 2.23.5 with the impure M-structure:



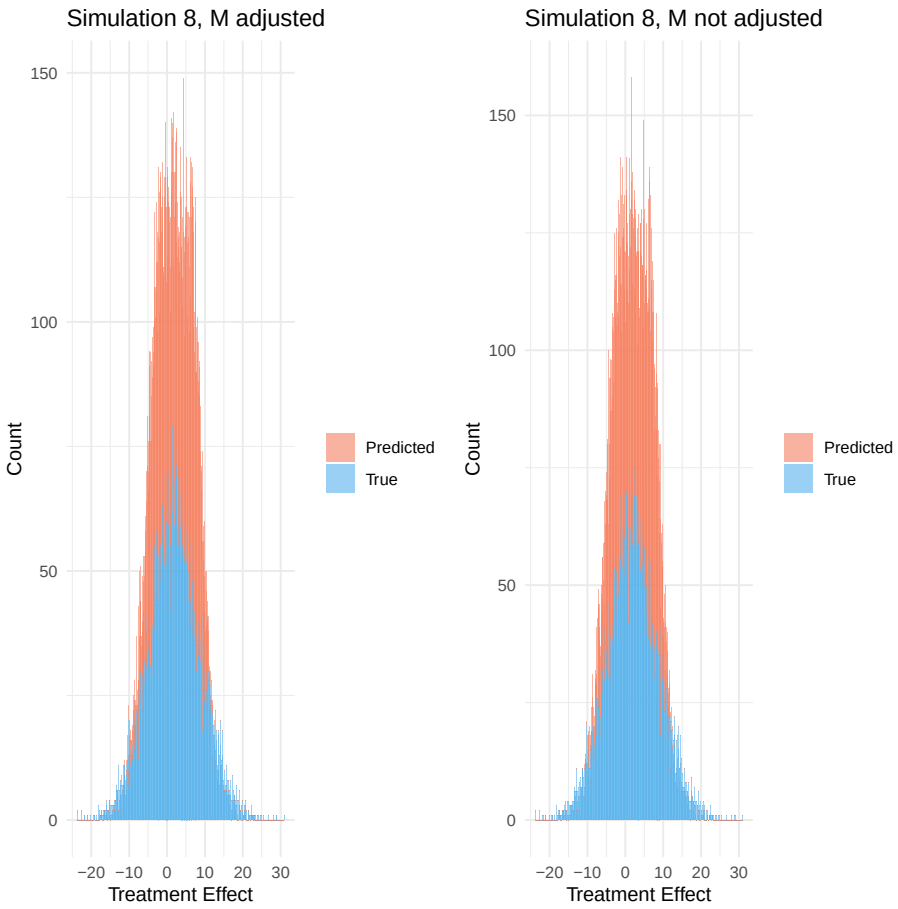


Figure 2.13: Simulation 8 predicted and true treatment effects.

	RMSE
M adjusted	3.201408
M not adjusted	3.218593

	Coverage
M adjusted	0.6700
M not adjusted	0.6671

Table 8: RMSEs and coverages in simulation 8.

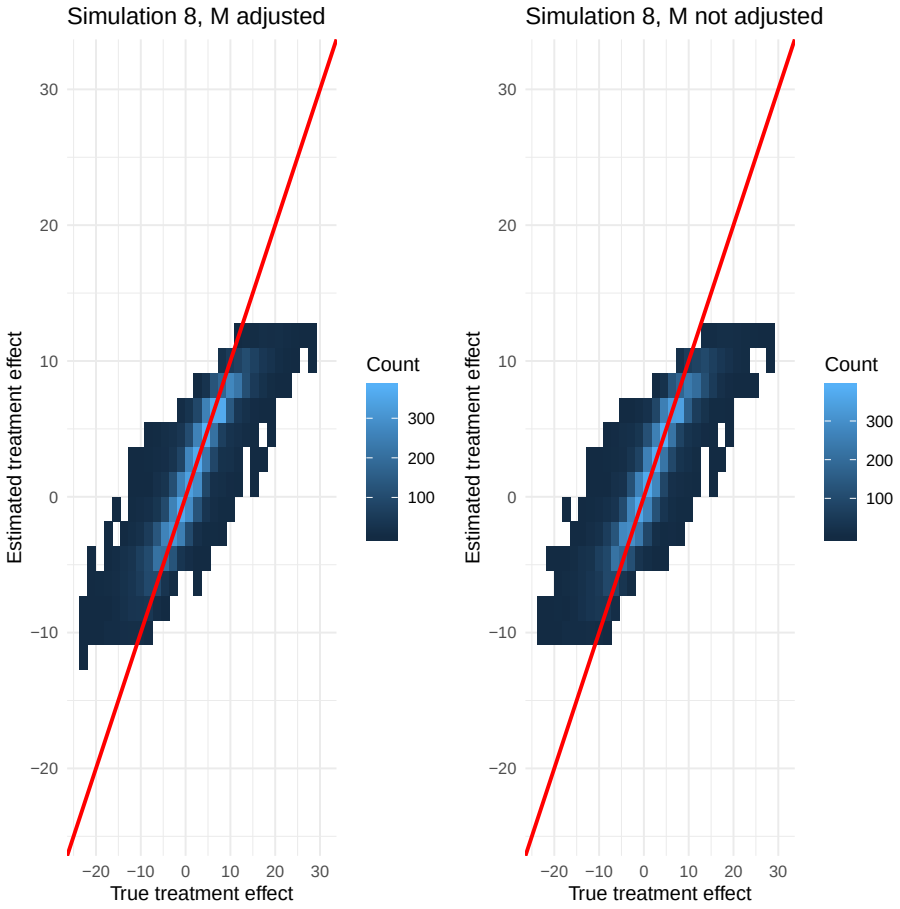


Figure 2.14: Simulation 8 predicted and true treatment effects.

$$f_W(Z_1, Z_2, C_1, \ldots, C_{10}, U_1, U_2, U_W) = \frac{1}{1 + \exp\left\{-\left(\sum_{j=1}^{2} \gamma_{Z_j} Z_j + \sum_{l=1}^{10} \gamma_{C_l} C_l + \sum_{m=1}^{2} \gamma_{U_m} U_m + U_Y\right)\right\}}$$

$$f_M = \delta_1 U_1 + \delta_2 U_2$$

$$f_Y(W, X_1, \ldots, X_{10}, C_1, \ldots, C_{10}, U_2, U_Y) = \tau(x_1, x_2, x_3, c_1) \cdot W + \sum_{k=1}^{10} \beta_{X_k} X_k + \sum_{l=1}^{10} \beta_{C_l} C_l + \beta_{U_2} U_2 + U_Y \quad , \text{ where}$$

$$\begin{cases} U_Y & \sim \text{Norm}(\mu_Y(0), \sigma_Y^2) \\ U_W & \sim \text{Norm}(0, \sigma_W^2) \\ Z_1, Z_2 & \sim \text{Bern}\left(\frac{1}{2}\right) \\ X_K & \sim \text{Norm}(\mu_K, \sigma_K^2) \text{ when } k \text{ is odd and} \\ X_K & \sim \text{Bern}\left(\frac{1}{2}\right) \text{ when } k \text{ is even} \\ C_L & \sim \text{Norm}(\mu_L, \sigma_L^2) \text{ when } l \text{ is odd and} \\ C_L & \sim \text{Bern}\left(\frac{1}{2}\right) \text{ when } l \text{ is even} \\ U_1, U_2 & \sim \text{Bern}\left(\frac{1}{2}\right) \end{cases}$$

and $\tau(x_1, x_2, x_3, c_1) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 x_3 + \alpha_3 c_1$.

The two estimated models are one with the M in the adjusted set $\mathbb{Z}_{\text{IncM}}$ and the other without, $\mathbb{Z}_{\text{ExcM}}$. The effect of adjusting the collider is once again moderate, but in contrasts to the earlier, in terms of RMSE and coverage, the model with $\mathbb{Z}_{\text{ExcM}}$ is superior (table 9). Still the distribution of the estimated treatment effects is more centered than the distribution of the true treatment effects, and the coverage rates are below 70 percent.

3 Discussion

In the simulations 2.23.1–2.23.8, more variables implicated improvement in terms of the RMSE and the coverage. The only exception to this was simulation 2.23.9 in which the model not including the collider M performed better than the model including M. However, the difference in the performances between the two models was moderate. According to the results of the simulation study, a practical recommendation would be to include as many relevant pretreated, non-instrumental variables in the model as possible. This is in line with the PO framework literature, as discussed in subsection 2.6. Another reason why one should avoid leaving pre-treated variables not-adjusted is that they may have influence in treatment effect: DAGs have no information which variables are important with respect to heterogeneous treatment effects. In his article, Imbens [2004] mentions two points to be considered in covariate selection: the first one is to avoid adjusting what should not be adjusted, namely post-treatment

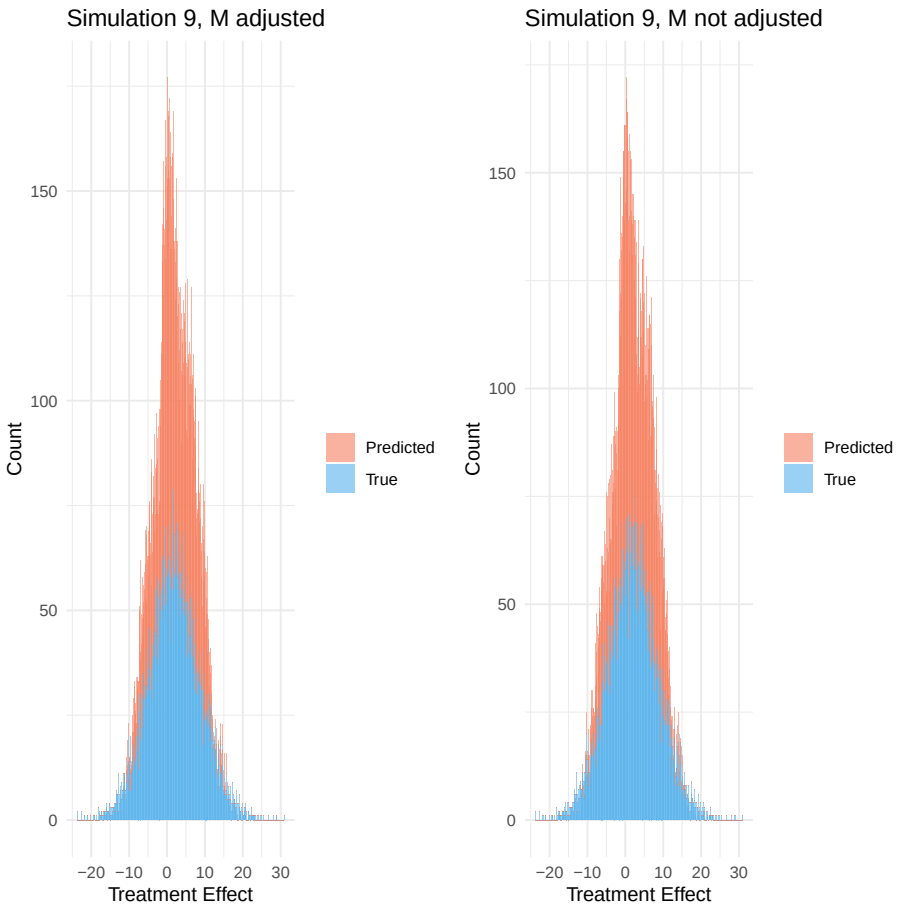


Figure 2.15: Simulation 9 predicted and true treatment effects.

	RMSE
M adjusted	3.291505
M not adjusted	3.279869

	Coverage
M adjusted	0.6589
M not adjusted	0.6664

Table 9: RMSEs and coverages in simulation 9.

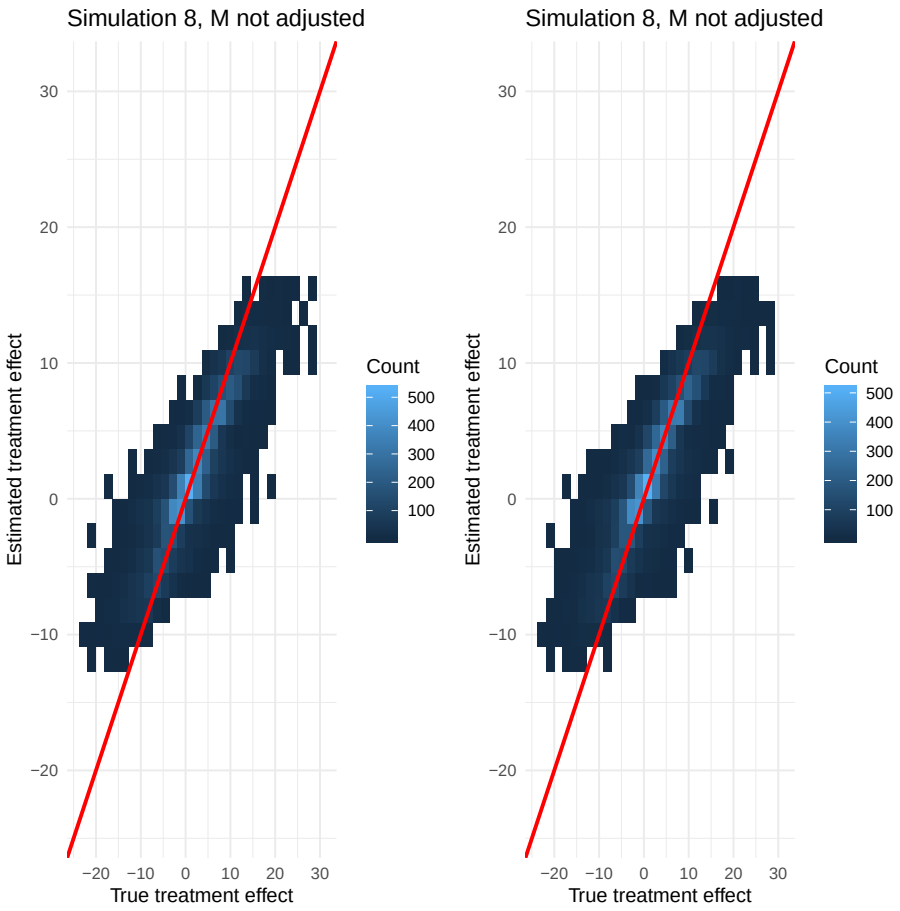


Figure 2.16: Simulation 9 predicted and true treatment effects.

variables, and the second one is to consider the use of variables that are weakly correlated with the treatment indicator and the outcomes. The latter point is not as critical with the causal forest and other machine learning algorithms that that determine the sample weights in the data-driven way as with traditional statistical methods such as matching or regression [Wager and Athey, 2018].

The first point that Imbens [2004] mentioned in the covariate selection was to avoid covariates that should not be adjusted. There exists a consensus for not-adjusting post-treatment variables, and thus the post-treatment case was not considered in the simulation study. In contrast, some other causal structures of the pre-treated variables, such as the M-structure, have been more controversial. An interesting result in the simulation was that adjusting the collider M seems to improve the results marginally in cases 2.23.4 and 2.23.8, when in theory adjusting for a collider in the M-structure should introduce bias in the estimates by opening a back-door path between the assignment and the output. As presented earlier in subsection 2.16, this has been a controversial topic in earlier literature. Another interesting result was that when in simulation 2.23.5 adjusting the collider M gave a better performance, the opposite was true with the corresponding simulation 2.23.9. Nevertheless, it is important to notice that the results of the models with adjusted sets $\mathbb{Z}_{\text{IncM}}$ and $\mathbb{Z}_{\text{ExcM}}$ are nearly identical in scenarios 2.23.4, 2.23.5, 2.23.8 and 2.23.9, and thus the relative orders may be coincidental.

Earlier studies [Greenland, 2003, Liu et al., 2012] have stated that the collider bias is usually relatively low, which can offer one part of the explanation for the results: even if there were some M-bias, it may get hidden beneath the randomness of the estimation process. One point is that the causal forest method is not strictly a pure stratification method, but rather a weighting process, where the weights are matching functions that are constructed by prioritizing important variables with respect to the causal effects. Thus, it is not totally clear how relevant the earlier discussion of the collider-stratification bias is with the causal forest algorithm. As mentioned in subsection 2.23.3, variable M has no significant importance according to the splittings and thus probably has a marginal role in the weightings. Deeper discussion about the potential pros and cons for including the collider in the models is out of the scope of this thesis, but according to results of this thesis and earlier studies, no matter whether the effect positive or negative, the effect is probably relatively low.

The low coverage rates of the causal forest have been discussed earlier in literature [Wager and Athey, 2018, Wendling et al., 2018, Athey et al., 2019]. In the article written by Athey et al. [2019], the researchers found that the coverage rates decline considerably when the treatment effect function is dependent on a higher number of variables: the real coverage rates were close to the nominal coverage rates still with the bivariate treatment function, but significantly below that with a function with four variables. The other remarks were that the coverage results were better with larger sample sizes, when the ambient dimension p is smaller, when the true signal is sparser, and when the true signal is additive.

The generated simulations are designed to represent clearly the underlying

causal structures and thus are carried out with SCMs. The simulations are totally generic and there is no guarantee that they would be realistic: the covariate and treatment effect distributions can be unrealistic, the joint distribution of the variables is probably too simple et cetera. However, as Imbens [2004] states: *"...although it is useful to compare these techniques in such realistic settings, it is also important to compare them in an artificial environment where one is certain that the underlying assumptions are valid"*. There are many things that are not covered in the simulation study that would give beneficial knowledge in regard to the subject: one could use different sample sizes, or test how the parameter tuning in the causal forest would affect the results. Also different kinds of covariate distributions and causal functions would probably affect the results. It would also be interesting to see how other heterogeneous treatment effect estimation methods, such as well performed BART and causal boosting [Powers et al., 2017, Wendling et al., 2018], would be affected by the different causal structures.

In the simulation study of the thesis, the approach of the estimation process is presented only as simple as necessary for exploring the question of model selection. In real applications, there are many steps to consider that are not covered in this thesis, such as identification strategies, sensitive analysis and tests for accessing assumptions [e.g. Rosenbaum and Rubin, 1983b, Angrist et al., 1996, Angrist and Krueger, 1999, Spirtes et al., 2000, Imbens, 2003, Angrist, 2004, Imbens, 2004, Angrist and Pischke, 2008, Heckman, 2008, Imbens and Rubin, 2015]. One theoretically and practically interesting approach is to aggregate methods in heterogeneous treatment effect estimation: the practice of aggregating individual results over several separate algorithms has shown to be successful in practice [Mullainathan and Spiess, 2017] and can be applied to heterogeneous treatment effect estimation [Grimmer et al., 2017].

4 Conclusion

This thesis aimed to explore how the underlying causal structures should be taken into account in covariate selection when performing heterogeneous treatment effect estimation. The research question was studied by reviewing existing literature and by executing a simulation study. In the simulation study, the aim was to explore the performance of the causal forest algorithm with different covariate sets for the algorithm in each simulation. The simulations were generated with SCMs.

Earlier literature has provided different views on an approach in covariate selection. The PO framework literature has suggested to balance the distributions of the pre-treated covariates as far as possible between the treated and the control group, whereas in contrast the SCM literature has suggested to use prior knowledge of the causal structures and to systematically define the sufficient set of covariates to be adjusted. In some cases, these frameworks suggest different adjusting strategies, such as in the M-structure case that was presented in the thesis. Furthermore, the review of literature provided an overview of heteroge-

neous treatment effect estimation methods. The causal forest method, which was used in the simulation study as an estimation method, was covered in more detail.

According to the results of the simulation study, a practical recommendation would be to include all the observed relevant pre-treated covariates in the model. In every simulation except one, the model with the highest number of covariates performed the best with respect to RMSE and coverage. Surprisingly, this result even applied to the cases where the SCM literature suggests not to condition all the variables. Besides covariate selection, the orthogonalization method was also tested, but this did not improve the performance of the causal forest even in a confounded causal structure. One notable result was that the coverage rates of the causal forest estimates were significantly below their nominal rates if the heterogeneous treatment effect function depended on higher number (over three in the simulation study) of covariates.

References

A. Abadie, J. Angrist, and G. Imbens. Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings. Working papers 99–16, Massachusetts Institute of Technology (MIT), Department of Economics, 1999. URL `https://ideas.repec.org/p/mit/worpap/99-16.html`.

J. Aldrich et al. Cowles exogeneity and core exogeneity. Technical report, Economics Division, School of Social Sciences, University of Southampton, 1993.

I. Andrews and E. Oster. A simple approximation for evaluating external validity bias. *Economics Letters*, 178:58–62, 2019.

I. Andrews, M. Gentzkow, and J. M. Shapiro. Measuring the sensitivity of parameter estimates to estimation moments. *The Quarterly Journal of Economics*, 132(4):1553–1592, 2017.

J. D. Angrist. Lifetime earnings and the vietnam era draft lottery: evidence from social security administrative records. *The American Economic Review*, pages 313–336, 1990.

J. D. Angrist. Treatment effect heterogeneity in theory and practice. *The Economic Journal*, 114(494):C52–C83, 2004.

J. D. Angrist and A. B. Krueger. Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4):979–1014, 1991.

J. D. Angrist and A. B. Krueger. Empirical strategies in labor economics. In *Handbook of labor economics*, volume 3, pages 1277–1366. Elsevier, 1999.

J. D. Angrist and J. Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

J. D. Angrist and J. Pischke. The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, 24(2):3–30, 2010.

J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996.

D. R. Arday, P. Lapin, J. Chin, and J. A. Preston. Smoking patterns among seniors and the medicare stop smoking program. *Journal of the American Geriatrics Society*, 50(10):1689–1697, 2002.

O. Atan, J. Jordon, and M. van der Schaar. Deep-treat: Learning optimal personalized treatments from observational data using neural networks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 2071–2078, 2018. URL `https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16157`.

S. Athey and G. Imbens. The Econometrics of Randomized Experiments. Papers 1607.00698, arXiv.org, 2016a. URL `https://ideas.repec.org/p/arx/papers/1607.00698.html`.

S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016b.

S. Athey and G. W. Imbens. Machine learning methods that economists should know about. *Annual Review of Economics*, 11(1):685–725, 2019.

S. Athey and S. Wager. Estimating treatment effects with causal forests: An application, 2019.

S. Athey, D. Eckles, and G. W. Imbens. Exact p-values for network interference. *Journal of the American Statistical Association*, 113(521):230–240, 2018.

S. Athey, J. Tibshirani, S. Wager, et al. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

B. Auguie. *gridExtra: Miscellaneous Functions for "Grid" Graphics*, 2017. URL `https://CRAN.R-project.org/package=gridExtra`. R package version 2.3.

P. C. Austin. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in medicine*, 27(12):2037–2049, 2008.

P. C. Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3):399–424, 2011.

P. C. Austin and E. A. Stuart. Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine*, 34(28):3661–3679, 2015.

M. P. Bitler, J. B. Gelbach, and H. W. Hoynes. What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments. *American Economic Review*, 96(4):988–1012, 2006.

M. Bonetti and R. D. Gelber. Patterns of treatment effects in subsets of patients in clinical trials. *Biostatistics*, 5(3):465–481, 2004.

D. D. Boos and D. Nychka. *Rlab: Functions and Datasets Required for ST370 class*, 2012. URL `https://CRAN.R-project.org/package=Rlab`. R package version 2.15.1.

R. H. Bradley, L. Whiteside, D. J. Mundfrom, P. H. Casey, B. M. Caldwell, and K. Barrett. Impact of the infant health and development program (ihdp) on the home environments of infants born prematurely and with low birthweight. *Journal of educational psychology*, 86(4):531, 1994.

L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and regression trees*. CRC press, 1984.

J. Brooks-Gunn, F. Liaw, and P. K. Klebanov. Effects of early intervention on cognitive function of low birth weight preterm infants. *The Journal of pediatrics*, 120(3):350–359, 1992.

D. Card and A. B. Krueger. Minimum wages and employment: A case study of the fast food industry in new jersey and pennsylvania. Working Paper 4509, National Bureau of Economic Research, October 1993. URL `http://www.nber.org/papers/w4509`.

C. Carvalho, A. Feller, J. Murray, S. Woody, and D. Yeager. Assessing treatment effect variation in observational studies: Results from a data challenge, 2019.

K. Casey, R. Glennerster, and E. Miguel. Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan*. *The Quarterly Journal of Economics*, 127(4):1755–1812, 2012.

V. Chernozhukov and C. Hansen. An IV Model of Quantile Treatment Effects. *Econometrica*, 73(1):245–261, 2005.

D. Chetverikov, A. Santos, and A. M. Shaikh. The econometrics of shape restrictions. *Annual Review of Economics*, 10:31–63, 2018.

H. A. Chipman, E. I. George, and R. E. McCulloch. Bayesian cart model search. *Journal of the American Statistical Association*, 93(443):935–948, 1998.

S. Crystal, U. Sambamoorthi, J. T. Walkup, and A. Akıncıgil. Diagnosis and treatment of depression in the elderly medicare population: predictors, disparities, and trends. *Journal of the American Geriatrics Society*, 51(12): 1718–1728, 2003.

S. B. Dale and A. B. Krueger. Estimating the payoff to attending a more selective college: An application of selection on observables and unobservables. *The Quarterly Journal of Economics*, 117(4):1491–1527, 2002.

R. H. Dehejia and S. Wahba. Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics*, 84(1): 151–161, 2002.

P. Ding and L. W. Miratrix. To adjust or not to adjust? sensitivity analysis of m-bias and butterfly-bias. *Journal of Causal Inference*, 3(1):41–57, 2015.

R. F. Engle, D. F. Hendry, and J. Richard. Exogeneity. *Econometrica*, 51(2): 277–304, 1983.

A. D. Federman, A. S. Adams, D. Ross-Degnan, S. B. Soumerai, and J. Z. Ayanian. Supplemental insurance and use of effective cardiovascular drugs among elderly medicare beneficiaries with coronary heart disease. *Jama*, 286 (14):1732–1739, 2001.

Z. Fewell, S. G. Davey, and J. A. C. Sterne. The Impact of Residual and Unmeasured Confounding in Epidemiologic Studies: A Simulation Study. *American Journal of Epidemiology*, 166(6):646–655, 2007.

A. Finkelstein, S. Taubman, B. Wright, M. Bernstein, J. Gruber, J. P. Newhouse, H. Allen, K. Baicker, et al. The oregon health insurance experiment: Evidence from the first year. Working Paper 17190, National Bureau of Economic Research, 2011. URL `http://www.nber.org/papers/w17190`.

S. Firpo. Efficient Semiparametric Estimation of Quantile Treatment Effects. *Econometrica*, 75(1):259–276, January 2007.

R. A. Fisher. Statistical methods for research workers. *Oliver & Boyd, Edinburgh*, 1925.

E. S. Ford, A. H. Mokdad, W. H. Giles, and G. A. Mensah. Serum total cholesterol concentrations and awareness, treatment, and control of hypercholesterolemia among us adults: findings from the national health and nutrition examination survey, 1999 to 2000. *Circulation*, 107(17):2185–2189, 2003.

G. D. Friedman, I. Tekawa, M. Sadler, and S. Sidney. Smoking and mortality: the kaiser permanente experience. *Changes in cigarette-related disease risks and their implication for prevention and control. Rockville, MD, National Institutes of Health, National Cancer Institute*, 477:499, 1997.

J. H. Friedman. Multivariate adaptive regression splines. *The annals of statistics*, pages 1–67, 1991.

J. H. Friedman. Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378, 2002.

M. Gail and R. Simon. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, pages 361–372, 1985.

A. Gelman and G. Imbens. Why ask why? forward causal inference and reverse causal questions. Technical report, National Bureau of Economic Research, 2013.

M. M. Glymour. Using causal diagrams to understand common problems in social epidemiology. *Methods in social epidemiology*, pages 393–428, 2006.

D. Green and H. Kern. Modeling heterogeneous treatment effects in large-scale experiments using bayesian additive regression trees. *The Annual Summer Meeting of the Society of Political Methodology, Iowa City*, 2010.

S. Greenland. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*, 14(3):300–306, 2003.

J. Grimmer, S. Messing, and S. J. Westwood. Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Political Analysis*, 25(4):413–434, 2017.

T. Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica*, 11(1):1–12, 1943.

P. R. Hahn, J. S. Murray, and C. Carvalho. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects, 2017.

T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer, 2009. ISBN 9780387848846.

J. J. Heckman. Econometric causality. *International Statistical Review*, 76(1): 1–27, 2008.

J. J. Heckman and V. J. Hotz. Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training. *Journal of the American statistical Association*, 84(408):862–874, 1989.

J. J. Heckman and R. Robb. Alternative methods for evaluating the impact of interventions: An overview. *Journal of Econometrics*, 30(1):239–267, 1985.

J. L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

P. W. Holland. Causal inference, path analysis and recursive structural equations models. *ETS Research Report Series*, 1988(1):1–50, 1988.

J. V. Hotz, G. W. Imbens, and J. H. Mortimer. Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics*, 125(1-2):241–270, 2005.

K. Imai and M. Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013.

G. Imbens and D. B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge, 2015. ISBN 10.1017/CBO9781139025751.

G. W. Imbens. Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2):126–132, 2003.

G. W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, 86(1):4–29, 2004.

G. W. Imbens. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics, 2019.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: With Applications in R.* Springer Publishing Company, Incorporated, 2014. ISBN 1461471370, 9781461471370.

M. M. Joffe, T. R. Ten Have, H. I. Feldman, and S. E. Kimmel. Model selection, confounder control, and marginal structural models: review and new applications. *The American Statistician*, 58(4):272–279, 2004.

F. Johansson, S. Uri, and S. David. Learning representations for counterfactual inference. *International conference on machine learning*, pages 3020–3029, June 2016.

G. V. Kass. An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 29(2):119–127, 1980.

B. Kelcey and J. Carlisle. The threshold of embedded m collider bias and confounding bias. *Society for Research on Educational Effectiveness*, 2011.

A. Krueger and O. Ashenfelter. Estimates of the economic return to schooling from a new sample of twins. Technical report, National Bureau of Economic Research, 1992.

R. J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620, 1986.

E. E. Leamer. Let's take the con out of econometrics. *The American Economic Review*, 73(1):31–43, 1983.

E. E. Leamer. Vector autoregressions for causal inference? In *Carnegie-rochester conference series on Public Policy*, volume 22, pages 255–304. North-Holland, 1985.

R. I. Levy. The framingham study: The epidemiology of atherosclerotic disease. *JAMA*, 245(5):512–512, 1981.

J. A. List, A. M. Shaikh, and Y. Xu. Multiple hypothesis testing in experimental economics. Working Paper 21875, National Bureau of Economic Research, 2016. URL `http://www.nber.org/papers/w21875`.

W. Liu, M. A. Brookhart, S. Schneeweiss, X. Mi, and S. Setoguchi. Implications of m bias in epidemiologic studies: a simulation study. *American journal of epidemiology*, 176(10):938–948, 2012.

A. Luedtke and M. Laan. Super-learning of an optimal dynamic treatment rule. *The international journal of biostatistics*, 12:305–332, 2016.

J. Ma, N. L. Sehgal, J. Z. Ayanian, and R. S. Stafford. National trends in statin use by coronary heart disease risk category. *PLoS Medicine*, 2(5), 2005.

D. P. MacKinnon, G. Warsi, and J. H. Dwyer. A simulation study of mediated effect measures. *Multivariate behavioral research*, 30(1):41–62, 1995.

G. Maldonado and S. Greenland. Simulation Study of Confounder-Selection Strategies. *American Journal of Epidemiology*, 138(11):923–936, 1993.

R. L. Matzkin. Semiparametric estimation of monotone and concave utility functions for polychotomous choice models. *Econometrica: Journal of the Econometric Society*, pages 1315–1327, 1991.

N. T. McCall, P. Parks, K. Smith, G. Pope, and M. Griggs. The prevalence of major depression or dysthymia among aged medicare fee-for-service beneficiaries. *International journal of geriatric psychiatry*, 17(6):557–565, 2002.

A. Moneta. Graphical causal models and vars: an empirical assessment of the real business cycles hypothesis. *Empirical Economics*, 35(2):275–300, 2008.

S. Mullainathan and J. Spiess. Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2):87–106, 2017.

X. Nie and S. Wager. Quasi-oracle estimation of heterogeneous treatment effects, 2017.

J. Pearl. Bayesian netwcrks: A model cf self-activated memory for evidential reasoning. In *Proceedings of the 7th Conference of the Cognitive Science Society, University of California, Irvine, CA, USA*, pages 15–17, 1985.

J. Pearl. [bayesian analysis in expert systems]: comment: graphical models, causality and intervention. *Statistical Science*, 8(3):266–269, 1993.

J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

J. Pearl. Remarks on the method of propensity score. *Statistics in Medicine*, 28(9):1415–1416, 2009a.

J. Pearl. Myth, confusion, and science in causal analysis. 2009b.

J. Pearl. *Causality*. Cambridge university press, 2009c.

S. Powers, J. Qian, K. Jung, A. Schuler, N. H. Shah, T. Hastie, and R. Tibshirani. Some methods for heterogeneous treatment effect estimation in high-dimensions, 2017.

J. W. Pratt and R. Schlaifer. On the interpretation and observation of laws. *Journal of Econometrics*, 39(1-2):23–52, 1988.

J. R. Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

P. M. Robinson. Root- N-Consistent Semiparametric Regression. *Econometrica*, 56(4):931–954, 1988.

P. R. Rosenbaum. Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398):387–394, 1987.

P. R. Rosenbaum. Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(408):1024–1032, 1989.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983a.

P. R. Rosenbaum and D. B. Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(2):212–218, 1983b.

P. R. Rosenbaum et al. The role of a second control group in an observational study. *Statistical Science*, 2(3):292–306, 1987.

P. R. Rosenbaum et al. *Design of observational studies*, volume 10. Springer, 2010.

D. B. Rubin. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

D. B. Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1):34–58, 1978.

D. B. Rubin. *Matched sampling for causal effects*. Cambridge University Press, 2006.

D. B. Rubin. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine*, 26(1):20–36, 2007.

D. B. Rubin. Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? *Statistics in Medicine*, 28 (9):1420–1423, 2009.

W. R. Shadish and T. D. Cook. The renaissance of field experimentation in evaluating interventions. *Annual Review of Psychology*, 60(1):607–629, 2009.

U. Shalit, F. D. Johansson, and D. Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3076–3085. JMLR. org, 2017.

I. Shrier. Propensity scores [letter to the editor]. *Statistics in Medicine*, 27: 2740–41, 2008.

C. Sievert. *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC, 2020. ISBN 9781138331457. URL `https://plotly-r.com`.

A. Sjölander. Letter to the editor: Propensity scores and m-structures. *Statistics in medicine*, 28:1416–20; author reply 1420, 2009.

P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman. *Causation, prediction, and search*. MIT press, 2000.

J. Splawa-Neyman, D. M. Dabrowska, and T. P. Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 1923, 1990.

M. Taddy, M. Gardner, L. Chen, and D. Draper. Heterogeneous treatment effects in digital experimentation. *Unpublished Manuscript*, 12 2014.

L. Tian, A. Alizadeh, A. Gentles, and R. Tibshirani. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109, 2012.

J. Tibshirani, S. Athey, and S. Wager. *grf: Generalized Random Forests*, 2020. URL `https://CRAN.R-project.org/package=grf`. R package version 1.1.0.

H. R. Varian. Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2):3–28, 2014.

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

H. I. Weisberg and V. P. Pontes. Post hoc subgroups in clinical trials: Anathema or analytics? *Clinical Trials*, 12(4):357–364, 2015.

T. Wendling, K. Jung, A. Callahan, A. Schuler, N. H. Shah, and B. Gallego. Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Statistics in Medicine*, 37(23): 3309–3324, 2018.

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL `https://ggplot2.tidyverse.org`.

H. Wickham. *forcats: Tools for Working with Categorical Variables (Factors)*, 2019. URL `https://CRAN.R-project.org/package=forcats`. R package version 0.4.0.

H. Wickham and L. Henry. *tidyr: Tidy Messy Data*, 2019. URL `https://CRAN.R-project.org/package=tidyr`. R package version 1.0.0.

H. Wickham, R. FranÃ§ois, L. Henry, and K. MÃŒller. *dplyr: A Grammar of Data Manipulation*, 2019. URL `https://CRAN.R-project.org/package=dplyr`. R package version 0.8.3.

P. Wilson, R. D. Abbott, and W. P. Castelli. High density lipoprotein cholesterol and mortality. the framingham heart study. *Arteriosclerosis: An Official Journal of the American Heart Association, Inc.*, 8(6):737–741, 1988.

P. W. Wilson, R. B. D'Agostino, D. Levy, A. M. Belanger, H. Silbershatz, and W. B. Kannel. Prediciton of coronary heart disease using risk factor categories. *Circulation*, 97(18):1837–1847, 1998.

J. M. Wooldridge. *Introductory Econometrics: A Modern Approach*. Cengage Learning, 2015.

J. M. Wooldridge and G. M. Imbens. *Recent Developments in the Econometrics of program Evaluation*. National Bureau of Economic Research, 2008.

P. G. Wright. *Tariff on animal and vegetable oils*. Macmillan Company, New York, 1928.

L. R. Wulsin and B. M. Singal. Do depressive symptoms increase the risk for the onset of coronary disease? a systematic quantitative review. *Psychosomatic medicine*, 65(2):201–210, 2003.

D. S. Yeager, P. Hanselman, G. M. Walton, J. S. Murray, R. Crosnoe, C. Muller, E. Tipton, B. Schneider, C. S. Hulleman, C. P. Hinojosa, et al. A national experiment reveals where a growth mindset improves achievement. *Nature*, 573(7774):364–369, 2019.

Q. Zhao, D. S. Small, and A. Ertefaie. Selective inference for effect modification via the lasso, 2017.

Appendices

Table 10: Abbreviations

ATE	Average treatment effect
CATE	Conditional average treatment effect
DAG	Directed acyclic graph
IPTW	Inverse probability of treatment weighting
MSE	Mean squared error
PO	Potential outcomes
RCT	Randomized controlled trial
RMSE	Root mean squared error
SCM	Structural causal models
SSLN	Strong law of large numbers

A Unbiasedness of the ATE Estimator

Proof. By defining

$$D_i = W_i - \frac{N_t}{N} = \begin{cases} \frac{N_c}{N} & \text{if } W_i = 1 \\ -\frac{N_t}{N} & \text{if } W_i = 0 \end{cases}, \text{ so that } \mathbb{E}[D_i] = 0,$$

the expression 2.5 can be rewritten in the following way:

$$\hat{\tau} = \tau + \frac{1}{N}\sum_{i=1}^{N} D_i[\frac{N}{N_t}Y_i(1) - \frac{N}{N_c}Y_i(0)]$$

Then it can be shown that $\hat{\tau}$ is an unbiased estimator for τ when the assumption of a constant causal effect holds:

$$\begin{aligned}\mathbb{E}[\hat{\tau}] &= \mathbb{E}[\tau + \frac{1}{N}\sum_{i=1}^{N} D_i(\frac{N}{N_t}Y_i(1) - \frac{N}{N_c}Y_i(0))] \\ &= \mathbb{E}[\tau] \\ &= \tau\end{aligned}$$

□

B Variance of the ATE Estimator

It can be shown [e.g. Imbens and Rubin, 2015], that the sample variance of $\hat{\tau}$ is the following:

$$\mathbb{V}[\hat{\tau}] = \frac{S_c^2}{N_c} + \frac{S_t^2}{N_t} - \frac{S_{tc}^2}{N} \quad , \text{ where} \tag{B.1}$$

the sample variances of $Y(0)$ and $Y(1)$ are

$$S_c^2 = \frac{1}{N-1}\sum_{i=1}^{N}[Y_i(0) - \bar{Y}(0)]^2 \text{ and } S_t^2 = \frac{1}{N-1}\sum_{i=1}^{N}[Y_i(1) - \bar{Y}(1)]^2 \tag{B.2}$$

The variance of unit-level treatment effects in the sample is the following:

$$S_{tc}^2 = \frac{1}{N-1}\sum_{i=1}^{N}[Y_i(1) - Y_i(0) - (\bar{Y}(1) - \bar{Y}(0))]^2$$

The sample variances $B.2$ can be estimated in the following way:

$$S_c^2 = \frac{1}{N_c - 1}\sum_{i:W_i=0}^{N}[Y_i^{Obs} - \bar{Y}_c^{Obs}]^2$$

$$\text{and } S_t^2 = \frac{1}{N_t - 1}\sum_{i:W_i=1}^{N}[Y_i^{Obs} - \bar{Y}_t^{Obs}]^2$$

Due to the *"fundamental problem of causal inference"*, the term S_{tc}^2 cannot be estimated. As an estimator for $\mathbb{V}[\hat{\tau}]$ it is a general way to use the version of the $B.1$ where the term $\frac{S_{tc}^2}{N}$ is ignored:

$$\mathbb{V}_{\text{Neyman}}[\hat{\tau}] = \frac{S_c^2}{N_c} + \frac{S_t^2}{N_t} \tag{B.3}$$

This leads to an upwardly biased estimator for $\mathbb{V}[\hat{\tau}]$, implicating conservative confidence intervals. Still there are two cases when estimator $B.3$ is unbiased: if the treatment effect τ is constant or if the sample at hand is viewed as a random sample from an infinite population, meaning that $\mathbb{V}[\hat{\tau}]$ is unbiased for the variance of $\hat{\tau}$ viewed as an estimator of the population average treatment effect $\mathbb{E}[Y_i(1) - Y_i(0)]$, rather than as an estimator of the sample average treatment effect $\frac{1}{N}\sum_{i=1}^{N}[Y_i(1) - Y_i(0)]$. [e.g. Imbens and Rubin, 2015].

C Unconfoundedness Given a Balancing Score

Proof. Given the unconfoundedness 2.6, the expression 2.11 can be written in the following way:

$$\begin{aligned} &W_i \perp\!\!\!\perp Y_i(0), Y_i(0) | b(X_i) \\ &\quad \iff \mathbb{P}_w[W_i = 1 \mid Y_i(0), Y_i(1), b(X_i)] = \mathbb{P}_w[W_i = 1 \mid b(X_i)] \end{aligned} \tag{C.1}$$

Because $W_i \in \{0, 1\}$, the left-hand side of the C.1 can be written as a conditional expectation. By using iterated expectations:

$$\begin{aligned} &\mathbb{P}_w[W_i = 1 \mid Y_i(0), Y_i(1), b(X_i)] \\ &\quad = \mathbb{E}_w[W_i \mid Y_i(0), Y_i(1), b(X_i)] \\ &\quad = \mathbb{E}[\mathbb{E}_w[W_i \mid Y_i(0), Y_i(1), b(X_i), X_i] \mid Y_i(0), Y_i(1), b(X_i)] \end{aligned} \tag{C.2}$$

Given the unconfoundedness 2.6 and the definition of balancing scores 2.10, the C.2 can be showed to equal the right-hand side of the C.1, meaning that the unconfoundedness given a balancing score (2.11) holds:

$$\begin{aligned} &\mathbb{E}[\mathbb{E}_w[W_i \mid b(X_i)] \mid Y_i(0), Y_i(1), b(X_i)] \\ &\quad = \mathbb{E}[W_i \mid b(X_i)] \\ &\quad = \mathbb{P}_w[W_i = 1 \mid b(X_i)] \end{aligned}$$

□

D Balancing Property of the Propensity Score

Proof. We show that

$$W_i \perp\!\!\!\perp X_i \mid e(X_i) \tag{D.1}$$

The expression D.1 is equal with the following statement:

$$\mathbb{P}[W_i = 1 \mid X_i, e(X_i)] = \mathbb{P}[W_i = 1 \mid e(X_i)] \tag{D.2}$$

Because $e(X_i)$ is a function of X_i, the left-hand side of the equation D.2 can be written as $\mathbb{P}[W_i = 1 \mid X_i]$, which is the definition of the propensity score $e(X_i)$.

$$\mathbb{P}[W_i = 1 \mid X_i, e(X_i)] = e(X_i)$$

Given that $W_i \in \{0, 1\}$, the right-hand side of the D.2 can be written as a conditional expectation. By using iterated expectations:

$$\begin{aligned} &\mathbb{P}[W_i = 1 \mid e(X_i)] \\ &\quad = \mathbb{E}[W_i \mid e(X_i)] \\ &\quad = \mathbb{E}[\mathbb{E}[W_i \mid e(X_i), X_i] \mid e(X_i)] \\ &\quad = \mathbb{E}[e(X_i) \mid e(X_i)] \\ &\quad = e(W_i) \end{aligned}$$

□

E A Simple Algorithm for Estimating the Propensity Score

Algorithm 1 presented in Dehejia and Wahba [2002].

Data: Covariates X
Result: Estimated Propensity Score
Start with a parsimonious logit specification to estimate the score;
while *TRUE* **do**
  Sort data w.r.t. estimated propensity score (ASC);
  Stratify all observations of equal score range;
  Statistical test for all X, differences in means across treated and comparison units within each stratum are not significantly different from zero;
  if *X are balanced between treated and comparison observations for all strata* **then**
    BREAK;
  else if *X are not balanced for one stratum* **then**
    while *X are not in balance* **do**
      Divide the stratum into finer strata;
      Statistical test for all X;
    end
  else
    Modify the logit by adding interaction terms and/or higher-order terms of the covariate;
  end
end

Algorithm 1: Algorithm for estimating the propensity score

F Rules of do-calculus

These three rules of do-calculus, their explanations and their proofs are provided in Pearl [1995] (this theorem is directly cited). The following expression are used:

$G_{\overline{X}}$ The DAG obtained by deleting from G all arrows pointing to nodes in X.

$G_{\underline{X}}$ The DAG obtained by deleting from G all arrows emerging from nodes in X.

$G_{\overline{X}\underline{Z}}$ The DAG obtained by deleting from G all arrows emerging from nodes in X and all arrows pointing to nodes in X.

In addition, $\mathbb{P}(y \mid do(x), z) \triangleq \frac{\mathbb{P}(y,z|do(x))}{\mathbb{P}(z|do(x))}$ is the probability of $Y = y$ given that X is held constant at x and that $Z = z$ is observed.

Theorem. ***Rules of Do-Calculus***

Let G be the DAG associated with a causal model as defined in 2.19, and let $\mathbb{P}(\cdot)$ stand for the distribution induced by that model. For any disjoint subsets of variables $X, Y, Z,$ and W, we have the following rules:

Rule 1 *(Insertion/deletion of observations):*

$$\mathbb{P}(y \mid do(x), z, w) = \mathbb{P}(y \mid do(x), w) \qquad \text{if } (Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}}}$$

Rule 2 *(Action/observation exchange):*

$$\mathbb{P}(y \mid do(x), do(z), w) = \mathbb{P}(y \mid do(x), z, w) \qquad \text{if } (Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}\underline{Z}}}$$

Rule 3 *(Insertion/deletion of actions):*

$$\mathbb{P}(y \mid do(x), do(z), w) = \mathbb{P}(y \mid do(x), w) \qquad \text{if } (Y \perp Z \mid X, W)_{G_{\overline{X},\overline{Z(W)}}}$$

, *where $Z(W)$ is the set of Z-nodes that are not ancestors of any W-node in $G_{\overline{X}}$.*

G K-Fold Cross-Validation Algorithm

Data: Set $S =$ Covariates $X \cup$ Outcomes Y
Result: *Cross-validation estimate* CV_k
Divide the set S in k folds of approximately equal size;
for $i = 1$ *to* k **do**
  Select i fold as a validation set;
  Use other sets but i to fit the model;
  Fit the model;
  Use set i to count $\mathrm{MSE}_i = \frac{1}{n_1}\sum_{i=1}^{n_1}(Y_i - \hat{Y}_i)^2$;
end
$\mathrm{CV}_k = \frac{1}{k}\sum_{i=1}^{k}\mathrm{MSE}_i$;

Algorithm 2: K-Fold Cross-Validation

H Recursive Binary Split

Data: Data from the latest region R: Covariates $X \cup$ Outcomes Y
Result: Recursive binary split
Function `FnRecursive(`*Region* R`):`
  for $p = 1$ *to* P **do**
    for $t = 1$ *to* T **do**
      Divide R w.r.t. covariate X_p and a cutpoint t into two regions $R_1 = \{X \mid X_p < t\}$ and $R_2 = \{X \mid X_p \geq t\}$;
      Count $RSS = \sum_{i:X_i \in R_1(p,t)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:X_i \in R_2(p,t)} (y_i - \hat{y}_{R_2})^2$;
    end
  end
  Choose (p^*, t^*) s.t. equation RSS is minimized;
  Split the data w.r.t. (p^*, t^*) in regions R_1 and R_2;
  ;

Algorithm 3: Recursive binary split

I Boosting for Regression Trees

Based on the algorithm presented in James et al. [2014]:

Data: Training data: Covariates $X \cup$ Outputs Y
Result: The boosted model
Set $\hat{f}(x) = 0$ and $r_i = y_i$ for all i in the training set;
for $b = 1$ *to* B **do**
  Fit a tree $\hat{f}^b$ with d splits to training data (X, r);
  Update $\hat{f}^b$ by adding a shrunken version of the new tree: $\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x)$;
  Update residuals $r_i \leftarrow r_i - \lambda \hat{f}^b(x_i)$;
end
Output the boosted model $\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x_i)$;

Algorithm 4: Boosting for Regression Trees

J Coefficients in the Simulation Study

	Value
β_{X_1}	1.1846881
β_{X_2}	1.6731552
β_{X_3}	2.9565855
β_{X_4}	-1.8355449
β_{X_5}	2.8452661
β_{X_6}	-1.1567457
β_{X_7}	-1.5805033
β_{X_8}	1.4988092
β_{X_9}	1.3516565
$\beta_{X_{10}}$	2.8661819
β_{C_1}	2.9628036
β_{C_2}	0.8623509
β_{C_3}	-0.0993400
β_{C_4}	-1.1987373
β_{C_5}	-2.5123647
β_{C_6}	-1.7654443
β_{C_7}	0.2969675
β_{C_8}	0.9698760
β_{C_9}	0.9790413
$\beta_{C_{10}}$	0.3159898
β_{U_2}	5.0000000

	Value
γ_{Z_1}	0.3103331
γ_{Z_2}	-0.8675404
γ_{C_1}	0.2809952
γ_{C_2}	0.0842018
γ_{C_3}	-0.2081202
γ_{C_4}	-0.0699206
γ_{C_5}	-0.1153378
γ_{C_6}	-0.3488950
γ_{C_7}	-0.2784899
γ_{C_8}	0.3364730
γ_{C_9}	-0.3641025
$\gamma_{C_{10}}$	-0.1293683
γ_{U_1}	0.5000000

	Value
δ_1	5
δ_2	5

K R-Codes and used packages

All R-codes can be found from GitHub: https://github.com/juholaht/Thesis
The used packages are the following:

- `dplyr` [Wickham et al., 2019]
- `forcats` [Wickham, 2019]
- `ggplot2` [Wickham, 2016]
- `grf` [Tibshirani et al., 2020]
- `gridExtra` [Auguie, 2017]
- `plotly` [Sievert, 2020]
- `Rlab` [Boos and Nychka, 2012]
- `tidyr` [Wickham and Henry, 2019]